RICE UNIVERSITY

Computational and Theoretical Analysis of Influenza Virus Evolution and Immune System Dynamics

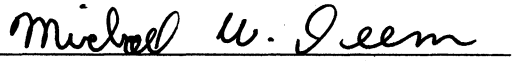
by

Keyao Pan

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

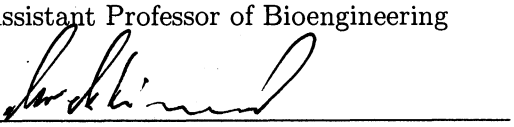
Doctor of Philosophy

Approved, Thesis Committee:



Michael W. Deem, Chair
John W. Cox Professor of Bioengineering
and Physics & Astronomy

Oleg A. Igoshin
Assistant Professor of Bioengineering



Marek Kimmel
Professor of Statistics

Houston, Texas

April, 2011

ABSTRACT

Computational and Theoretical Analysis of Influenza Virus Evolution and Immune System Dynamics

by

Keyao Pan

Influenza causes annual global epidemics and severe morbidity and mortality. The influenza virus evolves to escape from immune system antibodies that bind to it. The immune system produces influenza virus specific antibodies by VDJ recombination and somatic hypermutation. In this dissertation, we analyze the mechanism of influenza virus evolution and immune system dynamics using theoretical modeling and computational simulation.

The first half of this thesis discusses influenza virus evolution. The epidemiological data inspires a novel sequence-based antigenic distance measure for subtypes H1N1 and H3N2 virus, which are superior to the conventional measure using hemagglutination inhibition assay. Historical influenza sequences show that the selective pressure increases charge in immunodominant epitopes of the H3 hemagglutinin influenza protein. Statistical mechanics and high-performance computing technology predict fixation tendencies of the H3N2 influenza virus by free energy calculation. We introduce the notion of entropy from physics and informatics to identify the epitope regions of H1-subtype influenza A with application to vaccine efficacy. We also use entropy to quantify selection and diversity in viruses with application to the hemagglutinin of H3N2 influenza. Using the bacterial *E. coli* as a model, we show the evidence for

recombination contributing to the evolution of extended spectrum β-lactamases (ESBLs) in clinical isolates. A guinea pig experiment supports the discussion on influenza virus evolution.

The second half of the thesis discusses immune system dynamics. We design a two-scale model to describe correlation in B cell VDJ usage of zebrafish. We also introduce a dynamical system to model original antigenic sin in influenza.

This dissertation aims to help researchers understand the interaction between influenza virus and the immune system with a quantitative approach.

ACKNOWLEDGEMENTS

Probably I inherited the passion for mathematics, science, and engineering from my grandfather, Daoxu Pan, an outstanding electrical engineer who played a pivotal role in shaping my philosophy of nature when I was in my kindergarten. I am truly grateful for my parents, Chaoyi Pan and Yongfei Wang, for the education, support, and love they gave me. I also own my education and values to my grandmother, Saizhen Wang, and my maternal grandparents, Benjing Wang and Guiying Wu.

I would like to thank my advisor, Professor Michael W. Deem, for his indispensable inspiration, mentoring, and support throughout my graduate research. I would also like to thank my other committee members, Professors Oleg A. Igoshin and Marek Kimmel, for their critical remarks.

My graduate research was supported in part by DARPA grant HR 0011-09-1-0055, the National Science Foundation through TeraGrid resources provided by Purdue and Indiana Universities under grant number TG-MCA05S015, and a training fellowship from the Keck Center Nanobiology Training Program of the Gulf Coast Consortia (NIH Grant No. R90 DK071504).

Contents

Abstract ii

List of Illustrations xii

List of Tables xxix

1 Introduction 1

1.1 Influenza as a Global Health Threat 1

1.2 Biology of Influenza Virus 2

1.3 Vaccination and Antigenic Distance Measure 3

1.4 Prediction of Influenza Evolution 6

1.4.1 Using Shannon entropy and Relative Entropy 6

1.4.2 Using the Number of Charged Amino Acids 10

1.4.3 Using Free Energy Calculation 15

1.5 Modeling of Zebrafish Immune System 18

1.6 Original Antigenic Sin 21

1.7 Summary of Results 23

2 A Novel Sequence-Based Antigenic Distance Measure for H1N1, with Application to Vaccine Effectiveness and the Selection of Vaccine Strains 24

2.1 Introduction 24

2.2 Materials and Methods 27

2.2.1 Identities of Vaccine Strains and Dominant Circulating Strains 27

2.2.2 Estimation of Vaccine Effectiveness 27

2.2.3 Antigenic Distance Measured By Sequence Data 30
2.2.4 Antigenic Distance Measured by Hemagglutination Inhibition 31
2.3 Results 36
2.4 Discussion 44
2.4.1 Verification of the p_{epitope} Model 44
2.4.2 Comparison of H3N2 and H1N1 Vaccine Effectiveness and Evolution Rates 45
2.4.3 The p_{epitope} Model as a Supplement to HI Assay 48
2.5 Supplementary Materials 49
2.5.1 Humoral Immune System Plays a Major Role in Immunity to Influenza 49
2.5.2 Evaluation of Vaccine Effectiveness 50
2.5.3 Robustness of the p_{epitope} Model 52
2.5.4 Comparison of the p_{epitope} Model and the HI Assay for H3N2 Virus 56

3 Comment on Ndifon et al, "On the use of hemagglutination-inhibition for influenza surveillance: Surveillance data are predictive of influenza vaccine effectiveness" 57

4 The Epitope Regions of H1-Subtype Influenza A, with Application to Vaccine Efficacy 62

4.1 Introduction 62
4.2 Methods 64
4.2.1 Mapping the Epitope from H3 Hemagglutinin to H1 Hemagglutinin 64
4.2.2 Extension of mapped epitope using entropy method 65
4.2.3 Phylogenetic Tree and Its Root 66

4.2.4 Calculation of p_{epitope} 66
4.3 Results 68
4.3.1 Antigenic Distance 68
4.3.2 Epitope Identification 68
4.4 Discussion 73

5 Quantifying Selection and Diversity in Viruses by Entropy Methods, with Application to the Hemagglutinin of H3N2 Influenza 74

5.1 Introduction 75
5.2 Materials and Methods 80
5.2.1 Sequence Data 80
5.2.2 Histograms of 20 Amino Acids 80
5.2.3 Shannon Entropy as Diversity 81
5.2.4 Relative Entropy as Selection Pressure 82
5.3 Results 84
5.3.1 Correlation between Shannon entropy and relative entropy . . 85
5.3.2 Annual Virus Migration 89
5.3.3 Positions under Selection 91
5.3.4 Comparison of Different Regions 92
5.4 Discussion 96
5.5 Conclusion 99
5.6 Appendix: Monte Carlo Simulation of the Patterns of Selection and Diversity 100

6 Selective Pressure to Increase Charge in Immunodominant Epitopes of the H3 Hemagglutinin Influenza Protein 104

6.1 Introduction 105
6.2 Materials and Methods 110
6.2.1 Discrete–Time Markov Chain 116
6.2.2 Continuous–Time Markov Chain 118
6.2.3 Maximum Likelihood Estimation 119
6.2.4 Guinea Pig Animal Model 120
6.3 Results 121
6.3.1 Charge Increases in the Dominant Epitope 121
6.3.2 Evolution of Amino Acids in Epitopes of Hemagglutinin is Non–universal: Comparison with PAM Matrix 125
6.3.3 Guinea Pig Animal Model Verifies the Increase in Charge . . . 126
6.3.4 Partitioning the Amino Acids by Charge is Optimal 135
6.4 Discussion 138
6.4.1 Data Fitting and Verifying the RAMM Model 138
6.4.2 Model Reversibility 140
6.4.3 Fluctuation and Spatial Distribution of Charge 142
6.5 Conclusion 143
6.6 Supplementary Material 144
6.6.1 Fitting the Markov Models 144

7 Predicting Fixation Tendencies of the H3N2 Influenza Virus by Free Energy Calculation 146

7.1 Introduction 146
7.2 Materials and Methods 150
7.2.1 Scheme of the Free Energy Calculation 150
7.2.2 Einstein Crystal 151
7.2.3 Modified Hydrogen Atoms 158
7.2.4 Expressions of Free Energies 159

7.2.5 Implementation of Free Energy Calculation Algorithm 161
7.3 Results 165
7.3.1 Free Energy Landscape 165
7.3.2 Historical Substitutions in Epitope B 173
7.4 Discussion 177
7.4.1 Fitness of the Virus Strains 177
7.4.2 Selection in the Epitope 180
7.4.3 A Picture of the H3N2 Virus Evolution 181
7.4.4 Multiple Substitutions 184
7.4.5 Prediction of Future Virus Evolution 184
7.5 Conclusion 187

8 Evidence for Recombination Contributing to the Evolution of Extended Spectrum β-Lactamases (ESBLs) in Clinical Isolates 189

8.1 Introduction 190
8.2 Results 195
8.2.1 DnaSP 197
8.2.2 LDhat 198
8.2.3 Reticulate 200
8.2.4 The Max Chi Squared Algorithm and the Sawyer's Runs Test 202
8.2.5 PhylPro 203
8.2.6 The PHI Test 206
8.2.7 Linear and Logistic Regression Filters 207
8.3 Discussion 208
8.4 Materials and Methods 211

9 A Two-Scale Model for Correlation in B Cell VDJ Usage

of Zebrafish **215**

9.1 Introduction 216

9.2 Materials and methods 219

9.2.1 ODE model 220

9.2.2 Generalized NK model 222

9.2.3 Model characterization and verification 224

9.3 Results and discussion 225

9.3.1 Overview of the results 225

9.3.2 Primary and secondary immune response against one antigen . 227

9.3.3 Parameter sensitivity in the generalized NK model 230

9.3.4 Distribution of the energy change ΔU of a point mutation . . 232

9.3.5 Co-effect of multiple types of antigens 233

9.4 Conclusion and outlook 235

10 Understanding Original Antigenic Sin in Influenza with a Dynamical System **248**

10.1 Introduction 248

10.2 Materials and Methods 251

10.2.1 Characters of Influenza A Virus and Infection 251

10.2.2 Model Development and Description 253

10.2.3 Reduced Units and Parameter Estimation 256

10.3 Results 258

10.3.1 Time Courses of Infection and Recovery 258

10.3.2 A General Picture of Original Antigenic Sin 263

10.3.3 Mechanism of Original Antigenic Sin 266

10.3.4 Sensitivity Analysis 270

10.4 Discussion 271

11 Conclusion **275**

Bibliography **280**

Illustrations

1.1 The tertiary structure of the HA1 domain of H3 hemagglutinin (PDB code: 1HGF). The surface of HA1 facing outward is the exposed surface when the hemagglutinin trimer is formed. The other two HA1 domains (not shown) in the HA trimer are located at the back of the structure displayed here. The solid balls represent five epitopes. Color code: blue is epitope A, red is epitope B, cyan is epitope C, yellow is epitope D, and green is epitope E. 8

1.2 The scheme of the free energy calculation. The free energy difference of one substitution is calculated by $\Delta\Delta G = \Delta G_{43} - \Delta G_{21}$. State n, $n = 1$–4, is the real system. State na has the same configuration of atoms as state n except that all the hydrogen atoms have mass 16.000 amu. Compared to state na, state nb contains one additional Einstein crystal of product atoms ($n = 1, 3$) or reactant atoms ($n = 2, 4$). The mass of hydrogen atoms in state nb is also 16.000 amu. Free energy ΔG_{21b} and ΔG_{43b} are obtained by thermodynamic integration. 17

2.1 HA1 domain of the H1 hemagglutinin in the ribbon format (PDB code: 1RU7). Epitope A (blue), B (red), C (cyan), D (yellow), and E (red) are space filling. These five H1 epitopes are the analogs of the well-defined H3 epitopes [1]. 32

2.2 Vaccine effectiveness for influenza-like illness correlates with p_{epitope}, $R^2 = 0.68$ (solid line). Data from Table 8.2. The trend line quantifies vaccine effectiveness as a decreasing linear function of p_{epitope}. Vaccine effectiveness $= -1.19\ p_{\text{epitope}} + 0.53$. Also shown is the vaccine effectiveness to H3N2 (dashed line) [2]. 39

2.3 Vaccine effectiveness for influenza-like illness correlates with $p_{\text{all-epitope}}$ with $R^2 = 0.70$. Data from Table 8.2. The trend line quantifies vaccine effectiveness as a decreasing linear function of $p_{\text{all-epitope}}$. Vaccine effectiveness $= -4.16\ p_{\text{all-epitope}} + 0.54$. 40

2.4 Vaccine effectiveness for influenza-like illness correlates with p_{sequence} with $R^2 = 0.66$. Data from Table 8.2. The trend line quantifies vaccine effectiveness as a decreasing linear function of p_{sequence}. Vaccine effectiveness $= -7.37\ p_{\text{sequence}} + 0.54$. 41

2.5 The correlation with $R^2 = 0.53$ between vaccine effectiveness for influenza-like illness and d_1, the antigenic distance defined by HI assay using ferret antisera. Data from Table 8.2. The d_1 values were averaged if multiple HI assay experimental data were found. The trend line quantifies vaccine effectiveness as a decreasing linear function of d_1. Vaccine effectiveness $= -0.085\ d_1 + 0.50$. 42

2.6 The correlation with $R^2 = 0.46$ between vaccine effectiveness for influenza-like illness and d_2, the antigenic distance defined by HI assay using ferret antisera. Data from Table 8.2. The d_2 values were averaged if multiple HI assay experimental data were found. The trend line quantifies vaccine effectiveness as a decreasing linear function of d_2. Vaccine effectiveness $= -0.013\ d_2 + 0.51$. 43

2.7 The comparison between H3N2 (triangle up) and H1N1 (triangle down) in regard to the antigenic diversity, the evolutionary rate between 1980 and 2000 (left), the evolutionary rate between 2000 to 2007 (right), and the mutation rate on a short time scale without fixation. The antigenic diversity is measured with p_{epitope}, the unit of evolutionary rate is 10^{-3} nucleotide substitution/site/year, and the unit of mutation rate is 10^{-6} nucleotide substitution/site/day. 48

2.8 The linear regression with $R^2 = 0.0003$ of the residuals of H1N1 vaccine effectiveness versus year. Data from Table 1 and Figure 2 in the main text. The slope of the trend line is -0.0002/year. ANOVA test: H_0: slope $= 0$, $F = 0.0021$, and $p = 0.96$. The null hypothesis that these residuals are independent of time cannot be rejected. . . . 53

2.9 The linear regression with $R^2 = 0.018$ of the residuals of H3N2 vaccine effectiveness versus year. Data from [2]. The slope of the trend line is -0.0013/year. ANOVA test: H_0: slope $= 0$, $F = 0.32$, and $p = 0.58$. The null hypothesis that these residuals are independent of time cannot be rejected. 54

3.1 Vaccine efficacy versus the p_{epitope} or rAHM measures of antigenic distance. In inset are the data for which the vaccine and dominant circulating strain are distinct. 61

4.1 Phylogenetic tree of swine flu hemagglutinins deposited in NCBI and GISAID until 18 May 2009. For each tip containing over two strains, representative strains are marked. 67

4.2 Sequence entropy for the human strains of H1 (A/PR/8/1934 numbering, as in [3]). Positions belonging to predicted epitopes are color coded by the epitope identity. 70

4.3 Color-coded epitopes in the H1 structure (PDB code: 1RU7). 71

5.1 The tertiary structure of the HA1 domain of H3 hemagglutinin (PDB code: 1HGF). The surface of HA1 facing outward is the exposed surface when the hemagglutinin trimer is formed. The other two HA1 domains (not shown) in the HA trimer are located at the back of the structure displayed here. The solid balls represent five epitopes. Color code: blue is epitope A, red is epitope B, cyan is epitope C, yellow is epitope D, and green is epitope E. 77

5.2 Mean and standard error of relative entropy $S_{i+1,j}$ in each bin of Shannon entropy. Shannon entropy and relative entropy in each of the 329 positions and in each of the 17 seasons between 1992–1993 ($i = 0$) and 2008–2009 ($i = 16$) fall into one of the eight bins. The first bin with Shannon entropy less than 0.1 is discarded. Bins with larger Shannon entropy $D_{i,j}$ also have larger relative entropy $S_{i+1,j}$. Shannon entropy $D_{i,j}$ and relative entropy $S_{i+1,j}$ in iteration $i =$ 51–100 of the neutral evolution model are used to calculate mean and standard error of relative entropy in each bin of Shannon entropy distribution in the same way. No increasing trend is found. Error bar is one standard error. 87

5.3 Average Shannon entropy $\langle D\rangle_i$ versus average relative entropy $\langle S\rangle_{i+1}$ for each season between 1992–1993 ($i = 0$) and 2008–2009 ($i = 16$). For each season i, a set of amino acid positions j with Shannon entropy $D_{i,j}$ greater than 0.1 are chosen. For all the j in this set of positions, $\langle D\rangle_i$ is the average of the Shannon entropy $D_{i,j}$ values and $\langle S\rangle_{i+1}$ is the average of relative entropy $S_{i+1,j}$ values. Horizontal and vertical error bars are the standard errors of Shannon entropy and relative entropy, respectively. The solid line, $\langle S\rangle_{i+1} = 1.82\langle D\rangle_i - 0.23$, is a least squares fit of $\langle D\rangle_i$ to $\langle S\rangle_{i+1}$ ($i = 0, 2, \ldots, 16$). A strong correlation with $R^2 = 0.50$ exists between $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ excluding the point $(0.22, 1.38)$ with $N_i = 1$, which has a large standard error of the relative entropy $S_{i+1,j}$. Using the same method, $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ are calculated from a neutral evolution model, $i = 51$–100, and plotted. No visible correlation exists between $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ from the neutral evolution model. 88

5.4 (a) Average selection in each position quantified by relative entropy during the past 17 seasons from 1993–1994 to 2009–2010, calculated by $\bar{S}_j = \sum_{i=1}^{17} S_{i,j}/17$. The colors represent positions in epitopes A to E and positions outside the epitopes, as in Figure 5.1. (b) Number of seasons for each position when the relative entropy was greater than the threshold S_i^{thres}, i.e. the position was under selection. (c) Average diversity in each position quantified by Shannon entropy in the seasons from 1993–1994 to 2009–2010, calculated by $\bar{D}_j = \sum_{i=1}^{17} D_{i,j}/17$. (d) Distribution of the average selection in each position displayed in (a). (e) Distribution of the numbers of seasons under selection displayed in (b). (f) Distribution of the average diversity in each position shown in (c). 93

5.5 The results of the Monte Carlo simulation model containing epitopes A (blue) and B (red), and all the other positions. The model was simulated for 41 seasons. (a) Average selection in each position quantified by relative entropy calculated by $\bar{S}_j = \sum_{i=1}^{17} S_{i,j}/17$ in the last 17 seasons. The colors represent positions in epitopes A and B and positions outside the epitopes. (b) Number of seasons for each position when the relative entropy was greater than the threshold S_i^{thres}, i.e. the position was under selection. (c) Average diversity in each position quantified by Shannon entropy in the 17 seasons, calculated by $\bar{D}_j = \sum_{i=1}^{17} D_{i,j}/17$. (d) Distribution of the average selection in each position displayed in (a). (e) Distribution of the numbers of seasons under selection displayed in (b). (f) Distribution of the average diversity in each position shown in (c). 103

6.1 Number of charged amino acids for each year on the epitope A and epitope B of the dominant circulating strains from 1971 to 2003. . . . 111

6.2 Number of charged amino acids for each year on the epitope A and epitope B of the vaccine strains from 1971 to 2003. There were four consecutive intervals where epitope A or epitope B was dominant. . . 112

6.3 Average number of charged amino acids for each year on the epitope A and epitope B of the strains deposited in the GenBank database from 1971 to 2003. The curves in this figure were smoother than those in Figure 6.1 due to the averaging over database strains. The difference in the numbers of charged amino acids between this figure and Figure 6.1 is smaller than one for most years. 117

6.4 Number of charged amino acids for each year on the dominant epitopes of the circulating strains. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 are plotted. Since the estimated P_{c0} was close to the observed number, we used the observed numbers of charged amino acids in the first year of the interval to calculate P_{c0}. 122

6.5 Number of charged amino acids for each year on the dominant epitopes of the average of all the database strains in the same year. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 were plotted. We used the observed numbers of charged amino acids in the first year of the interval to calculate P_{c0}. Error bars for the circulating strains are one standard deviation calculated from all NCBI strains collected in that year. 124

6.6 Number of charged amino acids for each year on the dominant epitopes of the circulating strains. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 with PAM22 matrix were plotted. P_{c0} was fixed to the observed numbers of charged amino acids in the first year of the interval. The PAM22 matrix cannot reproduce the data. 127

6.7 Number of charged amino acids for each year on the dominant epitopes of the average of all the database strains in the same year. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 with PAM22 matrix were plotted. P_{c0} was fixed to the observed numbers of charged amino acids in the first year of the interval. The PAM22 matrix cannot reproduce the data. Error bars for the circulating strains are one standard deviation calculated from all NCBI strains collected in that year. 128

6.8 Comparison of charged residue changes between theoretical models and sequence data derived from Guinea pigs inoculated with the CDC A/Wyoming/2003 virus mixture. The RAMM and the PAM theoretical models were considered, as well as three data points from the Guinea pig infections: First point: Wyoming inoculum; Second point: progeny strains from infection of naïve Guinea pigs; and Third point: progeny strains from infection of previously infected Guinea pigs. The time of naïve and reinfection strains was estimated from the average number of amino acid mutations, counting both wild type and mutated strains, in the whole HA1 sequence. With the assumption derived from historical data that the annual mutation rate was 5.2 amino acids/HA1 sequence/year, we divided those average numbers of mutations by 5.2 to obtain the times. Error bars are one standard error. 132

6.9 Comparison of charged residue changes between theoretical models and sequence data derived from progeny virus isolated from Guinea pigs inoculated with the homogeneous WyB4 virus isolate. The number of predicted and observed charged residues were analyzed with the method used for the data in Figure 6.8. Error bars are one standard error. 133

7.1 The scheme of the free energy calculation. The free energy difference of one substitution is calculated by $\Delta\Delta G = \Delta G_{43} - \Delta G_{21}$. State n, $n = 1$–4, is the real system. State na has the same configuration of atoms as state n except that all the hydrogen atoms have mass 16.000 amu. Compared to state na, state nb contains one additional Einstein crystal of product atoms ($n = 1, 3$) or reactant atoms ($n = 2, 4$). The mass of hydrogen atoms in state nb is also 16.000 amu. Free energy ΔG_{21b} and ΔG_{43b} are obtained by thermodynamic integration. 152

7.2 The tertiary structure of the interface between the HA1 domain of H3 hemagglutinin monomer A/Aichi/2/1968 (bottom) and the antibody HC63 (top) (PDB code: 1KEN). Water molecules are not shown. Epitope B of the HA1 domain is located in two loops and one α-helix with the color scale modulated according to the expected free energy difference $\langle\Delta\Delta G\rangle_i$ of each site i in epitope B. The color scale ranges from red for the most negative $\langle\Delta\Delta G\rangle_i$ values to blue for the most positive $\langle\Delta\Delta G\rangle_i$ values. The sites i in epitope B with $\langle\Delta\Delta G\rangle_i$ near zero are colored white. The region outside epitope B is colored gray. The red site 128 is far from the antibody binding region and the red site 190 possessed the original amino acid Glu, which is a charged amino acid. It may explain why these two sites show negative $\langle\Delta\Delta G\rangle_i$ with large absolute values. 174

7.3 Two fixed substitutions G129A and E190D generated by Monte Carlo simulation of epitope B using equation 7.28. Also plotted are two historical fixed substitutions in epitope B: T155Y fixed between 1971 and 1973, and N188D fixed between 1970 and 1973. The frequency data of historical substitutions are from Shih et al. [4]. The origin of time axis is 1968. One thousand generation of the H3N2 virus is approximately one year. Figure 7.3(a) Substitution G129A causing the free energy difference $\Delta\Delta G = 3.33 \pm 0.29$ kcal/mol is fixed by the simulation. The rank of the free energy difference of G129A is 12 in 19 possible substitutions in site 129. Figure 7.3(b) Substitution E190D with $\Delta\Delta G = 18.75 \pm 0.32$ kcal/mol. The rank is 1 in 19 possible substitutions in site 190. 185

7.4 Two fixed substitutions N188D and V196D generated by Monte Carlo simulation of epitope B using equation 7.29. Two historical fixed substitutions T155Y and N188D are also plotted, and data are from Shih et al. [4]. Figure 7.4(a) Substitution N188D causing the free energy difference $\Delta\Delta G = 19.77 \pm 0.37$ kcal/mol is fixed by the simulation. The rank of the free energy difference of N188D is 1 in 19 possible substitutions in site 188. Figure 7.4(b) Substitution V196D with $\Delta\Delta G = 9.25 \pm 0.34$ kcal/mol. The rank is 5 in 19 possible substitutions in site 196. The proportions of substituting amino acids are represented by different line types. 186

8.1 Reticulate output for TEM (a) and SHV (b) respectively. The locations of the polymorphic sites in the genes are labeled on the axes of the squares. The black cells indicate genetic patterns that cannot be explained by point mutation as the sole mechanism of evolution. . 201

8.2 PhylPro results for TEM (a) and SHV (b) respectively. Each black curve represents a phylogenetic profile generated by PhylPro [5]. The red curve is the mean of all the phylogenetic profiles. In independent evolutionary trajectories that would come about in a mutation-only evolution process, the mutations in each particular trajectory are strongly correlated to each other. Hence the graph for the correlation measure for a gene pool that has evolved under a mutation-only scenario would be away from the Correlation = 0 line. Recombination destroys this correlation between different polymorphic sites, causing the graph to significantly deviate from the Correlation = 1 line. The higher the recombination rate, the further the plot deviates from the Correlation = 1 line and moves towards the Correlation = 0 line. (a) PhylPro results for TEM are close to the Correlation = 0 line at many positions. These low correlations can come from high recombination rates or low selection pressure [6]. The selection of TEM is usually intensive in the presence of antibiotics. Therefore, these low correlations suggests high recombination rates of TEM. (b) Similarly, PhylPro results for SHV are close to the Correlation = 0 line at all the positions, suggesting even higher recombination rates of SHV. . . 205

9.1 Illustration of the two-scale model of zebrafish immune response. The timeline arrow represents the zebrafish life history from 0 (hatch) to 180 days. To the right of the timeline is the first scale. Antigen challenged the zebrafish at several random timepoints. Each challenge led to a primary or secondary immune response. To the left of the timeline arrow is the second scale of the model. The flow chart describes the generalized NK model. Distinct secondary structures represented by squares with different colors were first built by minimizing the energy using Metropolis Monte Carlo method. These secondary structures randomly recombined to form V, D, and J segments, which randomly recombined to form the V region of the IgM heavy chain. The V region underwent 30 rounds of mutation and selection in the primary immune response and another 30 rounds in the secondary immune response. In this figure, as an example, the generalized NK model describes the primary immune response against the second antigen this zebrafish met in its lifetime, which is represented by the blue star. 239

9.2 (a) The numbers of distinct genotypes and VDJ recombination in the primary (generation 1–30) and secondary (generation 31–60) immune responses against one antigen involving $N_{\text{size}} = 2000$ naïve B cells. The number of VDJ recombination decreased much faster than that of B cell genotypes. In most cases all the B cells showed 1–2 VDJ recombinations at the end of the primary immune response and one VDJ recombination at the end of the secondary immune response. (b) Probability distribution at generation 60 of the rank of probabilities of $39 \times 5 \times 5 = 975$ VDJ recombination in $N_{\text{size}} = 2000$ naïve B cells reacting to one type of antigen. This probability distribution used 1000 rank data generated by running the generalized model 1000 times. The naïve VDJ ranks fell into 10 bins with rank 1–100, 101–200, ..., 901–975. Bin 10 with rank 901–975 was empty. 240

9.3 Trajectories of correlation coefficients r between VDJ usage of B cells in two zebrafish reacting to one certain antigen. The simulation consisted of 1000 runs, each of which generated the correlation coefficient r between naïve B cells in generation 0 and their progenies in generation 1–30 in the primary immune response and in generation 31–60 in the secondary immune response. The first 100 out of 1000 trajectories are here plotted for clarity. 241

9.4 (a) Relationship between the number of B cells reacting to one antigen, defined as N_{size}, and the correlation data between two zebrafish. Measured by the generalized NK model, the correlation coefficient r between the VDJ usage in the mature B cells from the secondary immune response fell into three categories: identical ($r > 0.995$), distinct ($r < 0.1$), and unfixed VDJ usage, with probabilities p, q, and $1 - p - q$, respectively. The values of p, q, and $1 - p - q$ were plotted respectively as the functions of N_{size} ranging from 10^3 to 10^5. (b) The number of distinct genotypes in each generation of the B cells in primary immune response. The x- and y-axes are the same as those in figure 9.2(a). This diagram presents the dynamics of genotype numbers in three cases, $N_{\text{size}} = 1000$, $N_{\text{size}} = 2000$, and $N_{\text{size}} = 10000$, respectively. (c) Same as (b), except for the number of different VDJ recombinations. 242

9.5 The effect of the average number of point mutations in each generation, n_{mut}, and the proportion of B cells propagated to the next generation, p_{cut}, on the probability p that two zebrafish developed mature B cells with correlated VDJ recombination against the antigen recognized by both zebrafish at the end of the secondary immune response. Each point on the surfaces shows the value of p calculated from the generalized NK model as a function of n_{mut} and p_{cut}. Each of the four subfigures is shown for distinct numbers of antigen-specific B cells N_{size}: (a) $N_{\text{size}} = 1000$, (b) $N_{\text{size}} = 2000$, (c) $N_{\text{size}} = 5000$, and (d) $N_{\text{size}} = 10000$. 243

9.6 The maximum number of B cells with identical sequence at the end of the secondary immune response, $N_{\max}$, as a function of n_{mut} and p_{cut}. As in figure 9.5, N_{size} as the number of antigen-specific B cells has a constant value in each subfigure: (a) $N_{\text{size}} = 1000$, (b) $N_{\text{size}} = 2000$, (c) $N_{\text{size}} = 5000$, and (d) $N_{\text{size}} = 10000$. 244

9.7 (a) The histogram of the energy difference $\Delta U = \hat{U} - U$ associated with a point mutation in the generalized NK model, in which U and $\hat{U}$ are the energy values before and after the point mutation. We calculate U and $\hat{U}$ for 1000 B cells at 61 generations. The values of ΔU ranged from -1.78 to 3.69. The histogram was equally divided in the interval $(-2, 4)$ by 12 bins. The relative frequency of the mutations with $\Delta U < 0$ is 0.038. (b) The original energy, U, versus the energy difference, ΔU, during 20 runs of the generalized NK model. The horizontal dashed line is $\Delta U = 0$. The solid line is the trend line between U and ΔU fit through the data. On average U decreases with the generation. At generation 0, U is typically close to -10, and at generation 60, U is typically close to -25. 245

9.8 Dynamics of mature B cells on day 0–180 in one zebrafish reacting to all types of the antigen in the environment. The numbers of plasma cells and memory B cells are the sums of all types of plasma cells and memory B cells, respectively. Each zebrafish was challenged by 10 types of antigen, the inoculation time of which followed a Poisson process as described in the main text. The dynamics shown in this figure are from one trajectory of the Poisson challenge process. 246

9.9 Distribution of correlation coefficients between the VDJ usage in B cell repertoires in distinct zebrafish on day 180. Using the categorization scheme in [7], the correlation coefficients r fall into four bins: no correlation with $r < 0.1$, low correlation with $0.1 \leq r < 0.2$, moderate correlation with $0.2 \leq r < 0.5$, and high correlation with $r \leq 0.5$. The height of each bar quantifies the relative frequencies of the correlation coefficient data in each bin. (a) Correlation coefficient data in the experiment [7]. (b) A total of 6000 correlation coefficients were generated by the model. 247

10.1 Time courses of proportions of healthy cell (H), infected cell (I), and dead cell (D), virus load (V), concentration of naïve and memory antibodies (X_1 and X_2), and the affinity of naïve antibody (U_1). with the condition $U_2 = 10^{-3}$. Initially, $H(0) = 1$, $I(0) = D(0) = 0$, $V(0) = 0.01$, $X_1(0) = 10^{-4}$, $X_2(0) = 10^{-2}$, and $U_1(0) = 10^{-3}$. 260

10.2 Time courses of proportions of healthy cell (H), infected cell (I), and dead cell (D), virus load (V), concentration of naïve and memory antibodies (X_1 and X_2), and the affinity of naïve antibody (U_1). We set the condition $U_2 = 0.5$. The initial values of all the state variables equal those in Figure 10.1. 261

10.3 Trajectories of maximum virus loads, maximum percentages of dead cell, maximum immune effects of naïve and memory antibodies by equation 10.9 and 10.10, respectively, and trajectory of final avarage affinities of the antibodies by equation 10.11 in a series of independent simulations, as the function of the affinity of the memory antibodies (U_2). The dashed lines in a and b are the level of maximum virus loads and maximum percentages of dead cells with the lowest memory antibody affinity $U_2 = 10^{-3}$, respectively. 264

10.4 Trajectories of dead cell proportion (D) and virus load (V) with different (E). In each trajectory, $H(0) = 1$, $I(0) = 0$, $V(0) = 100$, and E is constant. Small E such as 0.1 is not able to remove all the virus when $t \to \infty$. Larger decay rates of both D and V are observed for larger E. 268

10.5 Trajectories of the controlling effect $E = U_1X_1 + U_2X_2$. In each trajectory, $H(0) = 1$, $I(0) = 0$, $V(0) = 0.01, X_1(0) = 10^{-4}, X_2(0) = 10^{-2}$, and $U_1(0) = 10^{-3}$. Each trajectory corresponds to one value of U_2. Small E such as 0.1 is not able to remove all the virus when $t \to \infty$. Larger decay rates of both D and V are observed for larger E. 269

10.6 Sensitivity analysis of parameters c_0 and s. (a) and (b) Maximum percentage of dead cells and average antibody affinity for different c_0. (c) and (d) Maximum percentage of dead cells and average antibody affinity for different s. Initial conditions and parameters other than c_0 and s are the same as those in Figure 10.3. 271

Tables

2.1 HI table with two strains and four HI titers. 33

2.2 Summary of results. Nine pairs of vaccine strains and dominant circulating strains in seven flu seasons in the Northern hemisphere were collected from literature. The quantities n_u, N_u, n_v, N_v, p_{epitope}, $p_{\text{all-epitope}}$, p_{sequence}, d_1, and d_2 are defined in Materials and Methods. Only those seasons when H1N1 virus was dominant in at least one country or region where vaccine effectiveness data were available were considered. Two different vaccines have occasionally been adopted in different geographic regions for the same season, in which case two sets of data were added in this table. An asterisk signifies that co-circulating H3N2 was also found in the same country or region in that season; however, the interference to the final result from H3N2 is expected to be small, and so the sets of data with a single asterisk were preserved. 34

3.1 Correlation of H3N2 Influenza A vaccine efficacy in humans with different measures of antigenic distance. 61

4.1 Amino acids in epitopes A, B, C, D and E of H1 (A/California/04/2009 numbering, modified from [3]). For A/PR/8/1934 numbering, amino acid numbers above 130 would have 1 subtracted from them. 72

5.1 The relative entropy between hemagglutinin sequences in the different regions in the current influenza season and sequences in these regions in the previous season. The minimum relative entropy in each column is marked in bold. The p values of the Wilcoxon signed-rank test between the minimum relative entropy and other relative entropy values in the same column are in the parentheses. Hemagglutinin sequences were collected from four geographic regions: China, Japan, the USA, and Europe. Seven seasons from 2001–2002 to 2007–2008 are used here. The relative entropy values listed in this table are averaged for all the sites and all the six pairs of consecutive seasons. These results imply that the H3N2 viruses in China, Japan, and the USA migrate from China, while the H3N2 virus in Europe migrates from USA. 90

5.2 Amino acid positions j under selection. To be included, the positions must be under selection, $S_{i,j} > S_i^{\text{thres}}$, in greater than two seasons. . . 94

5.3 Annual selection, fraction of positions under selection, and diversity in epitopes A to E, positions not in any of the epitopes, and the whole HA1 sequence. 95

6.1 Epitope A and B of dominant circulating strains. Two numbers of charged amino acids in each epitope in each year are presented; the first one is calculated from the dominant circulating strain, and the second one is the simple arithmetic average of all strains collected in that year and deposited in GenBank. 114

6.2 Sequence analysis of progeny virus isolated from nasal washes of infected Guinea pigs. CDC virus designates mixture of Wyoming HA sequence variants contained in initial virus stock. WyB4 virus designates purified stock from predominant isolate of CDC virus. Immune status refers to whether the animals were naïve, i.e. neither previously infected nor immunized with purified HA protein, or immune. Number of sequences refers to the number of HA genes examined in progeny virus isolated from nasal washes. 134

6.3 R^2 values for amino acid categories and epitope B, the dominant epitope, using dominant circulating strains in 1972–1987 from which the model is trained. Seven categories and four combinations of categories are presented here. These four combinations involving charged and hydrophobic amino acids are chosen as the supplement to the seven categories, because charged and hydrophobic amino acids, especially charged ones, are critical in protein-protein interaction and evolution [8, 9, 10, 11, 12]. 136

6.4 R^2 values for amino acid categories and combinations of epitopes involving epitope B, the dominant epitope, using dominant circulating strains in 1972–1987 from which the model is trained. Residues outside an epitope are denoted by O. The dominant circulating strains in the time span of 1972–1987 have epitope B as the dominant epitope. The H3N2 virus emerged in 1968, therefore less adaptation to host immune system was developed compared with other time span. The previous discuss indicates that epitope B had the immunodominance in this period of time, and this table shows the limited contribution of amino acids outside the dominant epitope to the pattern of evolution. 137

7.1 Summary of the calculated free energy differences $\Delta\Delta G$ in each amino acid site in epitope B from the wildtype amino acid to all 20 amino acids. The standard errors are also listed. The free energy difference and its standard error of the substitution from the wildtype amino acid to itself are both zero. The units of free energy differences and their standard errors are kcal/mol. 166

7.2 The rank of the average binding free energy difference of the single substitution from alanine to another amino acid over all the 21 amino acid sites in epitope B of hemagglutinin trimer. The rank correlates with the charge and the size of amino acid, and it is relatively uncorrelated to the hydrophobicity. Here we applied classifications of RasMol for the biochemical properties of the 20 amino acids [13]. The relative frequencies of 20 amino acids were counted from the H3 sequences in NCBI database from 1968 to 2009. 170

7.3 The rank of the average free energy difference $\langle \Delta\Delta G \rangle_i$ generated by a substitution in each amino acid site i of epitope B. 172

7.4 Substitutions occurred in epitope B of the hemagglutinin A/Aichi/2/1968 (H3N2) as of 1975. Also listed are the time when the substitutions were observed, and the free energy differences with standard errors. In each site of epitope B, all the 20 amino acid were sorted in the descending order by the free energy differences introduced by a substitution from the wildtype amino acid to 20 amino acids. The ranks of the substituting amino acid and the wildtype amino acid in each substituted site are listed in the column Rank (substituting) and Rank (WT), respectively. 176

7.5 Substitutions occurred in epitope B of H3 hemagglutinin between the vaccine strain and the dominant circulating strain in each season in which the H3N2 subtype was dominant. The free energy difference with standard error of each substitution is obtained using the free energy landscape in Table 7.1. The ranks of free energy differences sorted in the descending order are listed in column Rank (vaccine) and in column Rank (circulating) for the amino acids in the vaccine strain and the dominant circulating strain, respectively. 178

8.1 Detection of recombination in TEM and SHV 196
8.2 Summary of simulation characterization results showing the sensitivity and the specificity of an algorithm based on our evaluation using simulated datasets 196

10.1 Descriptions and units of the variables in the model. 256
10.2 Parameters extracted from experimental data. 257
10.3 Decay rates of different immune cells. 258

Chapter 1

Introduction

1.1 Influenza as a Global Health Threat

Influenza A virus circulates in the human population every year, typically causing 3–5 million severe illnesses and 250,000–500,000 fatalities all over the world [14]. Hemagglutinin (HA) and neuraminidase (NA) are two kinds of virus surface glycoproteins encoded by the influenza genome. The subtype of influenza is jointly determined by the type of hemagglutinin ranging from H1 to H16, and that of neuraminidase ranging from N1 to N9. On the surface of the virus membrane, HA exists as a cylindrical trimer containing three HA monomers, and each monomer comprises two domains, HA1 and HA2. Hemagglutinin is also a key factor in virus evolution, because it is the major target of antibodies, and HA escape mutation changes the antigenic character of the virus presented to the immune system. The H3N2 virus causes the largest fraction of influenza illness.

The recent outbreak of H1N1 pandemic flu has caused immediate international concern. From its earliest case in mid-March 2009 to mid-May 2009, 8000–9000 infections and 70–80 deaths were recorded in 40–50 countries and regions, and as of mid-May 2009, over 90% of infections and deaths were in Mexico and the USA [15]. Historically, three subtypes of influenza A virus have been able to circulate in the human population. The Spanish flu pandemic in 1918–20 was H1N1, which circulated in the world until 1957. H1N1 reappeared in 1977 and persists today [16].

The Asian flu pandemic in 1956–58 was H2N2, which spread widely in the human population during the time interval 1957–68 [17]. The Hong Kong flu pandemic in 1968–69 was H3N2, which has circulated in the human population as the dominant subtype until recently [17]. Other subtypes rarely infected humans, although cases of H5N1 and H9N2 have been reported.

1.2 Biology of Influenza Virus

Influenza A is a type of RNA virus, belonging to the family Orthomyxoviridae. A functional Influenza A particle consists of a spherical lipid bilayer shell, with 8 distinct RNA strains that encode 11 kinds of proteins. Two kinds of glycoproteins, hemagglutinin (HA) and neuraminidase (NA), adhere to the surface of the virus particle. HA contributes to the binding of the virus particle to the sialic acid on the surface of the target upper respiratory tract epithelial cells and facilitates subsequent fusion and entry [18, 19, 20]. NA is the key component that facilitates virus release from surface membrane of infected cells [21]. Nucleocapsid protein (NP), a nucleoprotein, encapsulates and transports viral RNA inside the host cell [22]. The matrix protein, M1, binds to the inner side of the viral lipid bilayer membrane and helps assemble virus particles inside host cells, and is the central component in budding of virus particles [21]. M2, a glycoprotein inside the lipid bilayer, serves as the ion channel adjusting the pH value inside the virus particle, uncoating the virus particle, and contributes to efficient virion replication [23]. Additionally, there are two non-structural proteins, namely NS1 and nuclear export protein (NEP, formerly named NS2), as well as four RNA polymerases replicating the virus RNA in the nucleus of the infected cell, namely PA, PB1, recently discovered PB1-F2 [24], and PB2. NS1 interferes with the cellular antivirus system [25], and NEP exports newly synthesized viral ribonu-

cleoprotein (RNP) complexes comprising viral RNA, NP, and PB1 from the nucleus [26].

The 2009 swine flu virus possesses H1 hemagglutinin (HA) and N1 neuraminidase (NA) on the surface of the virion, of which the hemagglutinin is the main target of host antibodies. The human immune system responds primarily to the five epitope regions of the hemagglutinin protein [27, 28]. Host antibodies bind to five epitopes in hemagglutinin and lead to high escape evolution rates of amino acids in the epitopes. An early identification of H1 epitopes was carried out by antibody mapping of the A/PR/8/1934 (H1) hemagglutinin, with an additional study of laboratory mutations [3]. However, these H1 epitopes contain far fewer amino acids than do the epitopes in H3 determined by modern methods [28] and are incomplete. Alignment of H1 strains in 1918–2009 indicates many mutation positions outside the originally identified epitopes. We here use sequence alignment and information entropy to complete the definition of H1 epitopes.

1.3 Vaccination and Antigenic Distance Measure

Vaccination is an effective way to reduce the influenza morbidity and mortality. The efficacies of influenza vaccines vary from year to year, in part due to different antigenic distances between the circulating influenza strains and the vaccine. Antigenic distance between a vaccine strain and a viral strain can be estimated by the number of mutations in the hemagglutinin sequence between the two strains [29, 30]. Ferret animal model studies are used to further refine the notion of antigenic distance. These methods correlate with epidemiological studies of vaccine efficacy in humans with $R^2 = 0.59$ and 0.43–0.57, respectively [2, 31]. By considering only those mutations that occur in the dominant epitope, the p_{epitope} theory provides a prediction

of vaccine efficacy that correlates with epidemiological studies of vaccine efficacy in humans with $R^2 = 0.81$.

Vaccine efficacy has a linear correlation with the antigenic distance between the vaccine strain and the circulating virus strain [2, 31]. Since p_{epitope} correlates well with influenza vaccine efficacy in humans, it can be used to estimate antigenic distance. For example, when pepitope is larger than 0.19, the vaccine no longer offers protection. This correlation can be used to find optimal strains for vaccine design with minimal antigenic distance from expected circulating strains. Here we calculate the antigenic distance for H1 influenza A and apply the method to evaluate efficacy of a candidate swine flu vaccine.

The annual trivalent vaccine for influenza contains one H3N2 strain, one H1N1 strain, and one influenza B strain. This vaccine is currently the primary tool to prevent influenza infection and to control influenza epidemics. Due to the fast evolution of the influenza virus, the components of the influenza vaccine are changed for many flu seasons. Even though the vaccine is usually redesigned to match closely the newly evolved influenza virus strains, there occasionally has been a suboptimal match between vaccine and virus. Partly for this reason, vaccine effectiveness has varied in different years. The desire to have a vaccine with high effectiveness makes the prediction of the circulating influenza strain for the next influenza season a key step in vaccine design. A goal of the WHO is to recommend vaccine strains for the next flu season that will have the smallest antigenic distances to the dominant circulating strains in the next flu season, which often means using the dominant circulating strains in the current flu season as a reference.

A variety of distance measures have been developed to evaluate the degree of match between the vaccine strain and the dominant circulating strain. The hemag-

glutinin protein (HA) of influenza is primarily focused upon for this distance calculation since hemagglutinin is the dominant antigen for protective human antibodies and exhibits the highest evolutionary rate among all the influenza genes [32]. A widely used definition of antigenic distance is calculated from hemagglutination inhibition data from ferret animal model studies. To compare a pair of strains, a 2-by-2 HI titer matrix is built, and the antigenic distance is extracted from this matrix. This distance can be further refined by a dimensional projection technique termed antigenic cartography [30]. The mathematical basis of antigenic cartography is the dimension reduction of the shape space in which each point represents an influenza virus strain and the distance between a pair of points represents the antigenic distance between the corresponding strains. Note that antigenic cartography does not yield the distance data itself, but assesses the distance between the given vaccine strain and dominant circulating strain by globally considering the effect of all the strains and the antigenic distances among them. In the original literature of antigenic cartography [30], hemagglutination inhibition data were the input of the antigenic cartography algorithm that obtains the final results of distances. Antigenic distances can also be defined by the amino acid sequences of the strains using computer-aided methods, in which the fraction of substituted amino acid in the dominant hemagglutinin epitope bound by antibody is defined by p_{epitope} as a sequence-based antigenic distance measure [2, 33, 1]. The amino acid sequences are downloaded from databases and processed to obtain these distance measures. The p_{epitope} sequence-based method has been shown to be an effective antigenic distance measure between two strains of H3N2 [34, 2, 33]. To be clear, antigenic distance is a quantity that should define difference of viral strains, as determined by the human immune system. Ferret HI data are not the only or even the best measure of antigenic distances.

The vaccine effectiveness, which varies from year to year, correlates with the antigenic distance between the vaccine strain and the dominant circulating strain. Thus the vaccine effectiveness can be predicted by calculating the antigenic distance. Such *a priori* estimation of the vaccine effectiveness guides health authorities to determine the appropriate strain for the vaccine component for the coming flu season. For H3N2 influenza, the p_{epitope} method offers a prediction of vaccine effectiveness that has a higher correlation coefficient with vaccine effectiveness in humans than do distances derived by other methods [2, 33].

1.4 Prediction of Influenza Evolution

1.4.1 Using Shannon entropy and Relative Entropy

A common strategy by which viruses evade pressure from the immune system is to evolve and change their antigenic profile. Viruses with a low evolutionary rate that infect only humans, such as the small pox virus [35], can be effectively controlled by vaccinating the human population. By contrast, viruses with a high evolutionary rate, such as HIV, hepatitis B, and influenza A, resist being eliminated by the immune system by generating a plethora of mutated virus particles and causing chronic or repeated infection. In this study, we take subtype H3N2 influenza A virus as a model evolving virus. H3 hemagglutinin is under selection by the immune response mainly on the five epitope regions in the HA1 domain [36], labeled epitopes A to E, as shown in Figure 5.1. The immune pressure and the escape mutation drive the evolution of the H3N2 virus. The underlying mutation rate of the HA gene is 1.6×10^{-5}/amino acid position/day [37], measured using the method modified from that in an earlier study on the HA mutation rate [38]. Note that the mutation rate does not

necessarily equal to the evolution rate, or the fixation rate. The mutation rate equals to the evolutionary rate only if the evolution is neutral. The non-neutrality of the HA evolution is shown in the Results section. Evolution of the hemagglutinin viral protein causes occasional mismatch between the virus and the vaccine and decreases vaccine effectiveness [2, 34]. As more amino acid substitutions are introduced into influenza sequences, the antigenic characteristics of influenza strains drift away [39], and influenza epidemic severity of subtype H1N1 [40] and subtype H3N2 [41] increases.

The H3N2 virus has a distinguished evolutionary history, largely affected by the immune pressure. The H3N2 virus emerged in the human population in 1968 and has been circulating in the population since 1968. The phylogenetic tree of H3 hemagglutinin since 1968 has a linear topology in which most sequences are close to the single trunk of the tree, and the lengths of the branches are short [42, 30, 43]. Historical hemagglutinin sequences fall into a series of clusters, each of which has similar genetic or antigenic features and circulates for 2–8 years before being replaced by the next cluster [44, 30]. The evolution of different amino acid positions of hemagglutinin shows a remarkable heterogeneity: a subset of positions undergo frequent change, while some positions are conserved [4]. This heterogeneity is quantified by the Shannon entropy at each position of the amino acid sequence of hemagglutinin [1]. Shannon entropy has been used to locate protein regions with high diversity, such as the antigen binding sites of T-cell receptors [45]. Shannon entropy has been used to identify antibody binding sites, or epitopes, which are under immune pressure and so are rapidly evolving [1]. The heterogeneity of amino acid substitution suggests that point mutations randomly occurring in distinct positions have different contributions to the virus fitness.

The selection pressure on the H3N2 virus to evolve is reflected in the difference

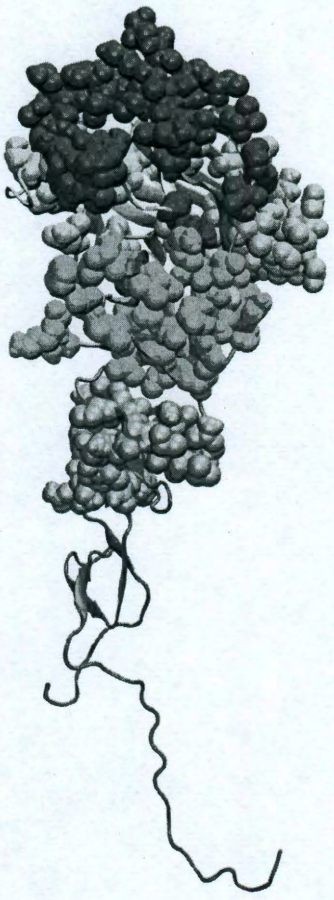


Figure 1.1 : The tertiary structure of the HA1 domain of H3 hemagglutinin (PDB code: 1HGF). The surface of HA1 facing outward is the exposed surface when the hemagglutinin trimer is formed. The other two HA1 domains (not shown) in the HA trimer are located at the back of the structure displayed here. The solid balls represent five epitopes. Color code: blue is epitope A, red is epitope B, cyan is epitope C, yellow is epitope D, and green is epitope E.

between the H3 hemagglutinin sequences in two consecutive seasons. We consider Northern hemisphere strains. When the epidemic initiates in a new season, we assume that each position of an HA sequence inherits the amino acid from a sequence of the previous season or has a different amino acid due to random mutation and selection. This assumption comes from the fact that the H3N2 virus circulating in each influenza season migrates from a certain geographic region in which the virus is preserved between two influenza seasons [43, 32]. In the absence of selection, the histogram of the 20 amino acid usage in one position in the current season is similar to that in the same position in the previous season except for changes due to the small random mutation rate. The difference between the two histograms beyond that expected due to mutation quantifies selection.

Synthesizing these factors, we introduce an entropy method to describe the evolution of influenza. The entropy method extracts an evolutionary pattern from aligned sequences. Shannon entropy quantifies the amount of sequence information in each position of aligned sequences [46, 47]. The sequence information reflects the variation, which is equivalently diversity, in each position, and so Shannon entropy has been used to measure the diversity in each position [48, 49, 50]. Shannon entropy has also been used to measure the structural conservation in the protein folding dynamics [51, 52]. See [53] for a detailed review of the applications of Shannon entropy. On the other hand, relative entropy measures gain of sequence information at each position and requires a background amino acid frequency distribution [54]. Relative entropy was also used as a sequence conservation measure to detect functional protein sites [55, 56]. Further, a dimension reduction technique using relative entropy has identified sectors in proteins [57, 56]. As an extension of these previous works, we apply Shannon entropy and relative entropy to jointly measure two quantities in

each position: sequence information in one season and gain of sequence information from one season to the next season. Simultaneous analysis of Shannon entropy and relative entropy sheds light on the evolutionary pattern of the H3N2 virus evolution when data from multiple seasons are available. In the HA1 domain, positions in the epitope regions have increased Shannon entropy, and this feature was applied to locate the epitopes of H1 hemagglutinin [1]. We here use Shannon entropy to quantify the virus diversity in each amino acid position in each season. The entropy relative to the previous season [58] is also used to analyze the evolution of the HA1 domain in one single season and to quantify the selection pressure on the virus in each amino acid position in each season. The selection and the virus diversity are two significant state variables determining the dynamics of evolution.

1.4.2 Using the Number of Charged Amino Acids

Influenza A virus causes annual global epidemics resulting in severe morbidity and mortality. The dominant circulating virus today is the H3N2 virus, which emerged in 1968 and is defined by two kinds of surface glycoproteins: H3 hemagglutinin and N2 neuraminidase. It is currently believed that hemagglutinin is relevant to virus attachment and entry into the cell, while neuraminidase facilitates virus release [20]. Hemagglutinin also plays a central role in the process of immuno escape, in which the antibodies mainly attack five epitopes, denoted as epitopes A–E, on the surface of the hemagglutinin protein [59, 28]. Because of antigenic changes through time, influenza vaccines are redesigned each year to provide improved protection against evolved circulating strains. The efficacy of the annual vaccine is variable due to the escape mutation of the influenza virus [29], especially mutation at the five epitopes on the hemagglutinin [2].

By analyzing the results of over 50 epidemiological studies of H3N2 influenza during the period 1968–2004, [2, 60, 31, 33] showed that the escape mutation of influenza A virus can be measured by p_{epitope}, the proportion of mutated amino acids in the dominant epitope of hemagglutinin, where the dominant epitope is defined as the epitope with the largest such proportion among the five epitopes. Compared with p_{sequence}, the proportion of mutated amino acids in the whole sequence of hemagglutinin, and the ferret antisera assays, p_{epitope} between vaccine strains and dominant circulating strains in the same flu season correlated better with the vaccine efficacies in the northern hemisphere [2]. Therefore p_{epitope} is an appropriate measurement for the antigenic distances between vaccine strains and dominant circulating strains. With the definition of the dominant epitope, the escape mutation at the dominant epitope induces the largest antigenic distance between vaccine strains and dominant strains, and endows the dominant epitope with the immunodominance.

In Gupta et al.'s model [2], which correlates well with vaccine efficacy in humans, every mutated amino acid is assigned the same weight. However, free energy calculations suggest that different amino acid substitutions have different contributions to the escape from the immune pressure. In general, the calculated differences in binding free energy $\Delta\Delta G = \Delta G_{\text{mutated}} - \Delta G_{\text{wildtype}}$ are different for different mutations, where $\Delta G_{\text{mutated}}$ and $\Delta G_{\text{wildtype}}$ denote binding free energy between two proteins one of which has and does not have a point mutation, respectively. For the experimentally measured difference in binding free energy $\Delta\Delta G$ between human growth hormone (hGH) and its first bound receptor (hGHbp), individual alanine substitutions of hydrophobic amino acids on the epitope of hGHbp induced the largest increase in $\Delta\Delta G$, followed by charged amino acids [8]. Nevertheless, we show that charged amino acids correlate more strongly with viral evolution. Nakajima [9] found that the majority

of escape mutations of H3 hemagglutinin of the strain A/Kamata/14/91 were the mutations that introduce charged amino acids, and that the frequency of selected mutations to charged residues was significantly higher than that expected by random chance. A related study by Smith et al. [30] found a similar over-representation of mutations to charged amino acids in the evolution of influenza.

We here consider the effect of different physical properties on the escape from immune pressure. We focus on the charged amino acids in the epitopes. These amino acids are strongly hydrophilic, and they reduce the tendency of antibodies to bind to hemagglutinin. Charged amino acids play a critical role in protein–protein interaction by creating salt bridges and salt bridge networks. Charged amino acids introduce specificity in binding [10]. Amino acid substitutions involving charged residues in the vicinity of receptor-binding region of HA affect the binding affinity between HA and its receptor [11]. The evolution of charged amino acids, therefore, may provide useful information on viral escape from antibody pressure. A discussion of charge evolution in proteins has been given by Leunissen [12], who observed a large variance in the evolutionary trends among different protein families. Here we use stochastic methods to model the evolution of charged amino acids on the epitopes of H3 hemagglutinin strains collected from humans since 1968.

The metaanalysis of 50 epidemiological human vaccine efficacy studies shows that the single dominant epitope is the critical region that determines the epidemiological vaccine efficacy [2]. There are five non-overlapping epitopes on the surface of H3 HA molecule, namely epitope A–E, to which different sets of antibodies bind. In each epitope, the p value is defined as the fraction of mutated amino acids [2]. The dominant epitope is defined as the epitope with the greatest p value. The greatest p value is p_{epitope}. Epidemiological data on the vaccine efficacies in 18 previous flu

seasons when H3N2 subtype was dominant were collected from approximately 50 studies [2]. The identities of the vaccine strains and dominant circulating strains were also obtained to calculate p_{epitope}. H3N2 vaccine efficacy correlates with p_{epitope} with $R^2 = 0.81$. This strong correlation shows that p_{epitope} defined by the single dominant epitope is a quantitative definition of antigenic distance. Importantly, the p_{epitope} calculated from the dominant epitope correlated better with vaccine efficacy than did antigenic distance including all HA amino acids [2].

The results of [61] show that subdominant epitopes are not the critical regions for vaccine efficacy. In an effort to improve the definition of antigenic distance, four modifications of the definition of antigenic distance were tested for their ability to improve the correlation with vaccine efficacy: 1) incorporating p values from subdominant epitopes, 2) distinguishing conservative and non-conservative amino acids mutations, 3) mutations in amino acids adjacent to the epitopes, and 4) the calculation of mutations in neuraminidase [61]. These four modifications of the definition of antigenic distance, including use of subdominant epitope p values, all failed to substantially improve the correlation with vaccine efficacy data in the years 1971–2004. These results motivate our focus on the dominant epitiope in the present analysis.

The reduced alphabet Markov model (RAMM) described in this thesis is an amino acid substitution model. It is built from the H3 hemagglutinin strains circulating in 1972 – 1987 when epitope B was the dominant epitope. This time span is shortly after the emergence of H3N2 virus in 1968, and this newly emerged virus subtype needed some time to adapt to host immune system, because emergence of new subtypes such as H2N2 in 1957 and H3N2 in 1968 went with Asian flu and Hong Kong flu outbreaks, indicating that subtypes like H3N2 were more virulent in the beginning and less adaptive to human. Further, the phylogenetic tree of H3N2 also shows

that H3N2 evolved faster at the very beginning than in the later stage [30]. So the pattern of evolution illustrates the escape mutation of the virus before substantial adaptation to host immune system was developed. Because the sequence database has been derived from patient samples and categorized by antigenic strain and date of collection, the human data do not necessarily reflect fixed variants, but, rather, snapshots along an evolutionary continuum.

Mutations of amino acids in different positions in the epitope are viewed as independent and identical Markov chains, whose parameters are the transition matrix $\mathbf{P}$ or the instantaneous rate matrix $\mathbf{Q}$. Markov models of protein evolution include the point accepted mutation (PAM) model [62] and the block substitution matrix (BLOSUM) model [63]. These models are derived by counting mutations in aligned amino acid sequences, and this approach provides the transition matrices $\mathbf{P}(t)$ of a Markov chain in a period of evolutionary time t. Adachi and Hasegawa [64] introduced the maximum likelihood method to estimate the elements of the transition matrix, and maximum likelihood was also employed to estimate the evolutionary time t when fixing the transition matrix $\mathbf{P}(t)$ [65]. The instantaneous rate matrix $\mathbf{Q}$ was calculated from the Laplace transform of $\mathbf{P}(t)$ [65, 66]. For a review of applications to 2000, see [67]. Some recent studies estimated the effect of possible multiple mutations at the same position within evolutionary time t [68, 69]. The instantaneous rate matrix has been estimated from observed frequencies of 20 amino acids [70].

The transition matrices $\mathbf{P}$ and the instantaneous rate matrices $\mathbf{Q}$ of most previous models are 20×20 matrices trained by analyzing databases with numerous alleles in many taxa. In the present case, the training data are limited, which can cause overfitting and introduce large errors to the fit model. One way to circumvent this difficulty is to decrease the number of parameters: 20 amino acids may be classified

into several groups with similar biophysical properties. Here we grouped 20 amino acids as 5 charged amino acids and 15 uncharged amino acids.

To test whether the mathematical models can be applied to data outside of the existing human data, we investigated their relevance to animal models of influenza infection. Guinea pigs have been shown to be infected with unadapted H3N2 strains. In most cases, the infection is limited to the upper respiratory tract, causes little apparent morbidity, and can be spread to cage mates via aerosol [71]. As part of our analysis of the p_{epitope} calculations, we developed additional evolutionary data derived through analysis of progeny viruses in animal model systems. Interestingly, the pattern of variations in the Guinea pig infection correlated well with the predictions of the model and with the human evolutionary data.

1.4.3 Using Free Energy Calculation

Influenza A virus causes annual global epidemics resulting in 5–15% of the population being infected, 3–5 million severe cases, and 250,000–500,000 fatalities [14]. The subtype of influenza A is determined by two surface glycoproteins—hemagglutinin (H) and neuraminidase (N). The H3N2 virus has been one of the dominant circulating subtypes since its emergence in 1968. The antibodies IgG and IgA are the major components of the immune system that control influenza infection, binding to the influenza hemagglutinin [72]. There are five epitopes at the antibody binding sites on the top of H3 hemagglutinin, namely epitopes A–E. The epitope bound most prolifically by antibody is defined as the dominant epitope, and it is central to the process of virus neutralization by antibody and virus escape substitution [2]. The cellular immune system, on the other hand, plays a relatively less recognized role in handling the invasive influenza virus [72]. The cellular system along with the

innate immune system exerts a somewhat more homogeneous immune reaction against genetically distinct influenza strains [73, 72].

Vaccination is currently the primary method to prevent and control an influenza epidemic in the human population [14]. Influenza vaccination raises the level of antibody specific for hemagglutinin and significantly enhances the binding affinity between antibody and hemagglutinin. Vaccine effectiveness depends on the antigenic distance between the hemagglutinin of the administered vaccine strain and that of the dominant circulating strain in the same season [2, 74]. Memory immune response from virus in previous seasons as well as vaccination in the current and previous seasons impose selective pressure on the current circulating virus to force it to evolve away from the virus strains recognized by memory antibodies that selectively bind to hemagglutinin.

As a result of the immune pressure and the escape evolution of the influenza virus, which is largely substitution in the dominant epitope of hemagglutinin, the influenza vaccine must be redesigned and administered each year, and the vaccine effectiveness has been suboptimal in some flu seasons [2, 33]. The escape evolution in the dominant epitope is at a higher rate than that in the amino acid sites outside the dominant epitope [42]. Sites in the dominant epitope also show higher Shannon entropy of the 20 amino acids than do those outside the dominant epitope [1]. High substitution rate and Shannon entropy in the dominant epitope of hemagglutinin suggest that the dominant epitope is under the strongest positive selection by human antibodies. The immune pressure against each genotype of the dominant epitope can be at least partially quantified by the binding constant between antibody and hemagglutinin.

The H3N2 virus and human immune system in this work are simplified to be a system consisting of the H3 hemagglutinin and the corresponding human antibody.

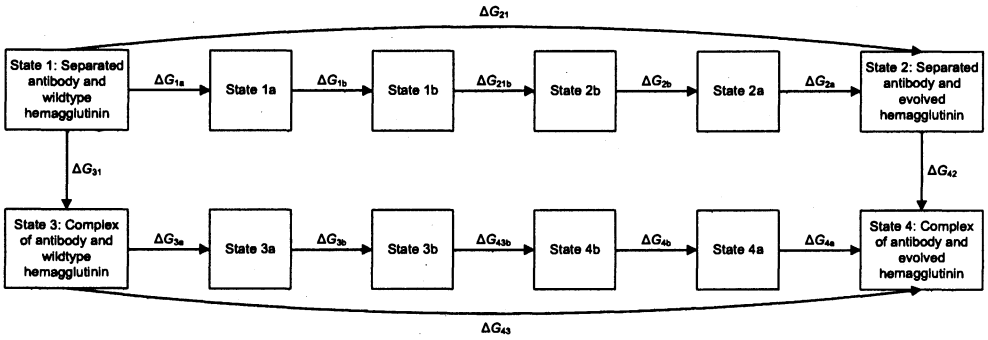


Figure 1.2 : The scheme of the free energy calculation. The free energy difference of one substitution is calculated by $\Delta\Delta G = \Delta G_{43} - \Delta G_{21}$. State n, n = 1–4, is the real system. State na has the same configuration of atoms as state n except that all the hydrogen atoms have mass 16.000 amu. Compared to state na, state nb contains one additional Einstein crystal of product atoms ($n = 1, 3$) or reactant atoms ($n = 2, 4$). The mass of hydrogen atoms in state nb is also 16.000 amu. Free energy $\Delta G_{21\mathrm{b}}$ and $\Delta G_{43\mathrm{b}}$ are obtained by thermodynamic integration.

Exposure by infection or vaccination produces an affinity-matured antibody with the binding constant to the corresponding hemagglutinin equal to 10^6–10^7 M^{-1}, while the binding constant of an antibody uncorrelated to the hemagglutinin is below 10^2 M^{-1} [72]. Escape substitutions may decrease the binding constant by changing the antibody binding free energy ΔG. Some substitutions decrease the antibody binding constant more than others and have higher probabilities to be fixed, because decrease in the antibody binding constant is favorable to the virus. Here we define the difference of antibody binding free energy as $\Delta\Delta G = \Delta G_{42} - \Delta G_{31}$ in which ΔG_{31} and ΔG_{42} are antibody-wildtype hemagglutinin binding free energy and antibody-evolved hemagglutinin binding free energy, respectively, as shown in 7.1. The fixation tendency of each substitution is a function of the difference of the antibody binding free energy [75] of the escape substitution.

Epitope A or B of the H3N2 virus was dominant in most influenza seasons [2]. Epitope B of the H3N2 virus was the dominant epitope presenting more substitutions than any other epitope in the recent years. Epitope B was also dominant in 1968 when H3N2 virus emerged. Thus during these periods of time, the substitutions in epitope B directly affect the antibody binding constant and reflect the direction of the virus escape substitution. To attain a global view of the effects of substitutions in epitope B, it is necessary to compute a matrix containing the differences of antibody binding free energy caused by each possible single substitution in epitope B. There are 21 amino acid sites in epitope B, and each residue in the wild type strain may substitute to any of the 19 different types to amino acid residues, hence we need to calculate a 19×21 matrix with 399 elements. Such a matrix is a free energy landscape quantifying the immune selection over each evolved influenza strain. In this free energy landscape, the virus tends to evolve to a position with low binding affinity of antibody to evade antibodies and reduce the immune pressure. Calculation of this landscape will enable us to study the mechanism of immune escape from a quantitative viewpoint, providing a criterion to describe and foresee the evolution of influenza virus.

1.5 Modeling of Zebrafish Immune System

B cell-mediated adaptive immunity exists in jawed animals [76]. B cells protect hosts by secreting antibodies that recognize and neutralize pathogens and foreign substances. Immunity generated by B cells is hence indispensable to the hosts' survival. The primary immune response occurs when a novel type of antigen is detected by the immune system. The antigen is processed and presented to naïve B cells, which mature in the germinal center. In the maturation process, the B cells acquire the ca-

pability to recognize and neutralize a specific antigen. First, a naïve B cell recombines one V gene segment, one D gene segment, and one J gene segment in the genome to create the nucleotide sequence encoding the antibody. Second, this nucleotide sequence undergoes multiple rounds of somatic hypermutation, and B cells with high affinity to the antigen are selected. The selected mature B cells differentiate into the antibody-secreting plasma cells or long-lived memory B cells that effectively activate the secondary immune response against the same antigen in the future. Janeway *et al.* provide a more detailed review of the B cell-mediated immune reaction [72]. Understanding the dynamics of and relationship between VDJ recombination and somatic hypermutation informs one about the central mechanism of B cell immunity.

Recent experimental studies provide information on the B cell maturation process in zebrafish. Zebrafish (*Danio rerio*) have been increasingly used as a model animal to study the immune system because experiments on zebrafish are easy to perform, zebrafish reproduce quickly, and zebrafish possess one of the most primitive adaptive immune systems, which is a model for the adaptive immune systems in humans and mice [77, 78, 79]. The genome of zebrafish contains 39 V gene segments, five D gene segments, and five J gene segments, which together encode the V region of immunoglobulin IgM heavy chain in zebrafish [80, 81]. High-throughput sequencing of the complete IgM repertoires in 14 six-month-old zebrafish revealed that one fish carries up to 5000–6000 distinct nucleotide sequence of IgM with $39 \times 5 \times 5 = 975$ VDJ recombinations [7]. VDJ usage, the probabilities that each of the 975 possible VDJ recombinations is used in the IgM repertoire, has a correlation coefficient between individual zebrafish up to $r = 0.75$ [7].

In the present study, we developed a two-scale model to illustrate that B cell maturation processes in distinct individuals, even though random, may converge to the

same VDJ recombination in the same environment of antigens. We use delay ordinary differential equations (ODEs) to simulate the immune response against multiple antigens circulating in the environment. We use the generalized NK model to describe B cell maturation processes against one antigen. The original NK model builds a random rugged energy landscape on which peptides evolve [82, 83]. The parameters of the NK model for short peptides have been fit to the observed data. Mora *et al.* fit a random energy model, similar to the NK model, with a large number, approximately 10^3, of parameters to experimentally measured probabilities of D gene segment usage in zebrafish [84]. As an extension of the original NK model, the generalized NK model takes into account the interaction between distinct subdomains of a protein and protein-protein interaction [85]. The generalized NK model can describe maturation of the whole V region of antibodies [34] and evolution of proteins in general [86]. In this study, we use the generalized NK model to analyze the convergence in the VDJ usage of immune response of two fish exposed to the same set of antigens. We extracted results of the generalized NK model to assign VDJ recombination to each type of mature B cells, the dynamics of which are solved by the ODE model. Correlation coefficients between the VDJ usage calculated from theory agree with experiment, in which most pairs of zebrafish had a weak correlation with $r \leq 0.2$ and a small fraction of pairs had $r > 0.5$ [7].

The two-scale model is motivated by the nature of immune response. The adaptive immune system receives different antigens at various time points. An antigen can initiate a primary or secondary immune response, depending on whether the infected individual has seen the antigen before. During the immune response B cells undergo rounds of somatic hypermutation and selection. The ODE system models at a mean field level the dynamics of B cell repertoires [87, 88]. The generalized NK

model computes the repertoire of B cells that respond and evolve in response to the antigen. The generalized NK model explicitly simulates the somatic hypermutation and selection of the B cell repertoire reacting to a specific type of antigen in one individual [85, 34]. We combined the ODE model and the generalized NK model to build the present two-scale model, a model that can zoom out to yield a global view of the dynamics of B cell repertoires and zoom in to focus on the somatic evolution of the B cells reacting to one type of antigen.

1.6 Original Antigenic Sin

Immune memory is mounted from previous infection or vaccination, stores the information for recognizing the corresponding antigen, and is activated during the future infection of the same type of pathogen. Long-term immune memory has been observed for various pathogens including smallpox [89], malaria [90], hepatitis B [91], dengue [92], and Influenza A [93]. This long-lasting effect prevents the reinfection by pathogens such as smallpox via recognizing and rapidly eliminating those reinfected pathogen particles. Smallpox virus, also called variola virus, propagates only in humans and has a relatively low mutation rate [35]. In contrast, Influenza A virus propagates in human, pigs, and aquatic birds, with a higher mutation rate that is approximately 2.0×10^{-6}/nucleotide/infectious cycle [37], or 1.6×10^{-5}/amino acid/day. Calculation of the binding free energy between human antibodies and circulating Influenza A strains shows that the virus mutates away from the genotypes that code for hemagglutinin proteins that are well recognized by the human immune system [94]. Thus for Influenza A, there is a significant antigenic distance between the circulating strain in a given year and the immune memory from previous years.

Original antigenic sin is the phenomenon in which prior exposure to an antigen

leads to a subsequent suboptimal immune response to a related antigen. In some years, when the antigenic distance between the vaccine strain and the circulating strains fell into a certain range, the existence of original antigenic sin frustrated the effort to decrease Influenza A infection rate by vaccination. Historical data of the Influenza A vaccine indicates that vaccine efficacy does not monotonically decrease with the distance between the vaccine strains and the circulating strains, but rather has a minimum at an intermediate level of antigenic distance [2]. Interestingly, since the efficacy of the vaccine with this intermediate antigenic distance from the circulating strains is lower than the case with larger distance, which is equivalent to unvaccinated people, original antigenic sin could make vaccinated people more susceptible to the virus than those who are unvaccinated.

One of the earliest works discussing the mechanism of original antigenic sin at the antibody level includes [34], which attributed original antigenic sin to the localization of subsequent immune response in the amino acid sequence space around the primary one. The affinity between an antibody and an antigen is given by the generalized NK model (GNK model) derived from the NK model originally introduced to model a rugged fitness landscape [82, 95] and evolution processes [96, 97, 98], to model the three-dimensional structure of protein molecules rather than peptides. The sequences of a group of antibodies for Influenza A are allowed to mutate freely and independently in the affinity landscape to maximize the individual affinities to the virus. B cells that make antibodies with highest affinities are expanded and propagated to the next round of the simulation. The mutation of the virus is captured by changing the fitness landscape. The final average affinity correlates well with the observed data in history [2].

1.7 Summary of Results

The remaining part of this thesis is organized as follows: Chapter 2 uses the p_{epitope} theory to predict vaccine effectiveness. Chapter 3 provides further evidence that the p_{epitope} theory is more accurate than the conventional ferret animal model. Chapter 4 identifies the epitope regions of H1 hemagglutinin using Shannon entropy. Chapter 5 estimates future H3N2 evolution and migration using currently available H3 hemagglutinin sequences. Chapter 6 shows that the number of charged amino acids increased in the dominant epitope B of the H3N2 virus since introduction in humans in 1968. Chapter 7 used free energy calculations with Einstein crystals as reference states to calculate the difference of antibody binding free energy ($\Delta\Delta G$) induced by amino acid substitution at each position in epitope B of the H3N2 influenza hemagglutinin. Chapter 8 shows evidence for recombination contributing to the evolution in clinical conditions. Chapter 9 builds a two-scale model in zebrafish and to explain the reported correlation between VDJ usage of B cell repertoires in individual zebrafish. Chapter 10 explains the mechanism of original antigenic sin with a dynamical system. Finally, chapter 11 summarizes this thesis.

Chapter 2

A Novel Sequence-Based Antigenic Distance Measure for H1N1, with Application to Vaccine Effectiveness and the Selection of Vaccine Strains

H1N1 influenza causes substantial seasonal illness and was the subtype of the 2009 influenza pandemic. Precise measures of antigenic distance between the vaccine and circulating virus strains help researchers design influenza vaccines with high vaccine effectiveness. We here introduce a sequence-based method to predict vaccine effectiveness in humans. Historical epidemiological data show that this sequence-based method is as predictive of vaccine effectiveness as hemagglutination inhibition (HI) assay data from ferret animal model studies. Interestingly, the expected vaccine effectiveness is greater against H1N1 than H3N2, suggesting a stronger immune response against H1N1 than H3N2. The evolution rate of hemagglutinin in H1N1 is also shown to be greater than that in H3N2, presumably due to greater immune selection pressure.

2.1 Introduction

The annual trivalent vaccine for influenza contains one H3N2 strain, one H1N1 strain, and one influenza B strain. This vaccine is currently the primary tool to prevent influenza infection and to control influenza epidemics. Due to the fast evolution of the influenza virus, the components of the influenza vaccine are changed for many flu seasons. Even though the vaccine is usually redesigned to match closely the

newly evolved influenza virus strains, there occasionally has been a suboptimal match between vaccine and virus. Partly for this reason, vaccine effectiveness has varied in different years. The desire to have a vaccine with high effectiveness makes the prediction of the circulating influenza strain for the next influenza season a key step in vaccine design. A goal of the WHO is to recommend vaccine strains for the next flu season that will have the smallest antigenic distances to the dominant circulating strains in the next flu season, which often means using the dominant circulating strains in the current flu season as a reference.

A variety of distance measures have been developed to evaluate the degree of match between the vaccine strain and the dominant circulating strain. The hemagglutinin protein (HA) of influenza is primarily focused upon for this distance calculation since hemagglutinin is the dominant antigen for protective human antibodies and exhibits the highest evolutionary rate among all the influenza genes [32]. A widely used definition of antigenic distance is calculated from hemagglutination inhibition data from ferret animal model studies. To compare a pair of strains, a 2-by-2 HI titer matrix is built, and the antigenic distance is extracted from this matrix. This distance can be further refined by a dimensional projection technique termed antigenic cartography [30]. The mathematical basis of antigenic cartography is the dimension reduction of the shape space in which each point represents an influenza virus strain and the distance between a pair of points represents the antigenic distance between the corresponding strains. Note that antigenic cartography does not yield the distance data itself, but assesses the distance between the given vaccine strain and dominant circulating strain by globally considering the effect of all the strains and the antigenic distances among them. In the original literature of antigenic cartography [30], hemagglutination inhibition data were the input of the antigenic cartography

algorithm that obtains the final results of distances. Antigenic distances can also be defined by the amino acid sequences of the strains using computer-aided methods, in which the fraction of substituted amino acid in the dominant hemagglutinin epitope bound by antibody is defined by p_{epitope} as a sequence-based antigenic distance measure [2, 33, 1]. The amino acid sequences are downloaded from databases and processed to obtain these distance measures. The p_{epitope} sequence-based method has been shown to be an effective antigenic distance measure between two strains of H3N2 [34, 2, 33]. To be clear, antigenic distance is a quantity that should define difference of viral strains, as determined by the human immune system. Ferret HI data are not the only or even the best measure of antigenic distances.

The vaccine effectiveness, which varies from year to year, correlates with the antigenic distance between the vaccine strain and the dominant circulating strain. Thus the vaccine effectiveness can be predicted by calculating the antigenic distance. Such *a priori* estimation of the vaccine effectiveness guides health authorities to determine the appropriate strain for the vaccine component for the coming flu season. For H3N2 influenza, the p_{epitope} method offers a prediction of vaccine effectiveness that has a higher correlation coefficient with vaccine effectiveness in humans than do distances derived by other methods [2, 33]. In this chapter, we develop the p_{epitope} method for H1N1 influenza. In Materials and Methods we describe the epidemiological data used to calculate vaccine effectiveness and the animal model or sequence data used to calculate antigenic distance. In Results we show the correlation of antigenic distance with vaccine effectiveness. We discuss the results in the Discussion.

2.2 Materials and Methods

2.2.1 Identities of Vaccine Strains and Dominant Circulating Strains

The vaccine strain selection by WHO in each year follows a standard procedure. The vaccine strains are reviewed every year and are usually changed every two to three years. We used the H1N1 vaccine strains and H1N1 dominant circulating strains in the epidemiological literature that provided vaccine effectiveness data used in this study.

2.2.2 Estimation of Vaccine Effectiveness

The H1N1 vaccine effectiveness is gathered from epidemiological literature regarding the influenza-like illness rate of unvaccinated (u) and vaccinated people (v). Vaccine effectiveness can be described by the following definition:

$$\text{vaccine effectiveness} = \frac{u - v}{u}. \tag{2.1}$$

To calculate vaccine effectiveness and its standard error, we let N_u and N_v denote the number of subjects in the unvaccinated and vaccinated group, n_u and n_v denote the number of illness in the unvaccinated and vaccinated group, respectively. The

values and the standard errors of u, v, and vaccine effectiveness are

$$u = n_u/N_u \tag{2.2}$$

$$v = n_v/N_v \tag{2.3}$$

$$\mathrm{VE} = \frac{u-v}{u} = \frac{n_u N_v - n_v N_u}{n_u N_v} \tag{2.4}$$

$$\sigma_u = \sqrt{\frac{u\,(1-u)}{N_u}} \tag{2.5}$$

$$\sigma_v = \sqrt{\frac{v\,(1-v)}{N_v}} \tag{2.6}$$

$$\sigma_{\mathrm{VE}} = \left(\frac{v}{u}\right)\sqrt{\left(\frac{\sigma_v}{v}\right)^2 + \left(\frac{\sigma_u}{u}\right)^2} = \sqrt{\left(\frac{1}{u}\right)^2 \sigma_v^2 + \left(\frac{v}{u^2}\right)^2 \sigma_u^2}. \tag{2.7}$$

If the vaccine effectiveness is averaged from N studies, $\sigma^2_{\mathrm{VE}} = \left(\sum_i \sigma^2_{\mathrm{VE}i}\right)/N^2$ where $\sigma_{\mathrm{VE}i}$ is the standard error of the i–th study.

Compared to H3N2, subtype H1N1 viruses were dominant in fewer years. Based on the proportions of samples of H3N2, H1N1, and influenza B collected in each year during 1977–2009, widespread H1N1 circulation was observed in approximately 10 seasons. Epidemiological studies on vaccine effectiveness were absent for some years when H1N1 circulated. Additionally, we used the criteria listed below to filter all available literature.

To ensure that the vaccine effectiveness we collected from the literature is for H1N1, the seasons and the geographic regions of the epidemiological studies in the literature were compared with the influenza activity information in WHO Weekly Epidemiological Records to confirm that those regions were dominated by H1N1 in those seasons. Subjects were restricted to 18–64 year old healthy adult humans to avoid effects of an underdeveloped immune system in children or of immunosenescence in senior people. If more than one measure of vaccine effectiveness was collected for

the same season, they were averaged to minimize the statistical noise.

In order to minimize the effect on vaccine effectiveness from co-circulating subtypes such as H3N2, only the epidemiological data collected in the regions and in the flu seasons in which the H1N1 subtype was dominant were applied to calculate the vaccine effectiveness in this study. The seasons in which the H1N1 subtype was dominant were reported by the literature on H1N1 vaccine effectiveness. The studies cited in Table 8.2 for the calculation of vaccine effectiveness gave the subtype of the predominant epidemic virus as well as of the virus sampled from the subjects with influenza-like illness (ILI). In addition, the dominance of H1N1 subtype is also available in the CDC Morbidity and Mortality Weekly Reports and the WHO Weekly Epidemiological Record. For the data in Table 8.2, the dominance of H1N1 subtype was shown in these references.

The vaccine effectiveness collected from various flu seasons and regions were measured with standard errors. Biases in the vaccine effectiveness are due to the complexity of the vaccine effectiveness measurement, including the character of the human population studied, such as age, immune history, and health condition; the influence of co-circulating H3N2 influenza strains; the character of the vaccine distributed, such as live attenuated virus vaccine, inactivated split-virus vaccine produced by virion disassembly, or subunit vaccine only containing hemagglutinin and neuraminidase; the method of epidemiological measurement of influenza infection, such as virus detection, confirmed symptomatic influenza, or influenza-like illness (ILI); the design of the experiment, such as natural infection or experimental challenge study; and the progression of the epidemic in the population under study. These biases are thus inevitable with current technology. Here, we applied the following methods to minimize biases in the vaccine effectiveness data. Subjects in the studies were confined

to 18–64 years old healthy adult humans to preclude the interference of the feeble immune system in children or in senior people, because variation in the capability of the immune system is a determinant of the vaccine effectiveness given the same pair of vaccine strain and dominant circulating strain. Only epidemiological studies in the season and the region in which H1N1 subtype was dominant were used to obtain the vaccine effectiveness data. The vaccine involved in the referred studies is an inactivated vaccine. Other types such as cold-adapted nasal spray vaccine were excluded. The epidemiological measurement of infection in all the referred studies used ILI as the criterion. Not all studies designed the experiment as a challenge study. We assume that the epidemic propagates in the population in a similar way in each season. These criteria are used to filter the available references and to obtain vaccine effectiveness data with minimum bias. The standard errors of the data are presented here. These criteria reduced the number of practical references for each season. Our metaanalysis considered 50 peer-reviewed papers, all we could find in the literature. We list the ones that satisfy our selection criteria for each of the years, typically 1–3 *per year.*

2.2.3 Antigenic Distance Measured By Sequence Data

Figure 2.1 shows the HA1 domain with five epitopes of the H1 subtype hemagglutinin. As the improvement of a previous definition of H1 epitopes [3], these five H1 epitopes are recognized by host antibodies and are identified by mapping the well-defined epitopes in H3 hemagglutinin [36, 28] to H1 hemagglutinin and using sequence entropy to find additional sites under selection [1].

The antigenic distance between the vaccine strain and the dominant circulating strain is the input for the vaccine effectiveness prediction. The fraction of mutated

amino acids in the epitope region of HA, or the p-value, is an antigenic distance measure to quantify the similarity between two strains [2]. One p-value is calculated for each H1 epitope

$$p\text{-value} = \frac{\text{number of mutations in the epitope}}{\text{number of amino acids in the epitope}}. \tag{2.8}$$

The p_{epitope} is defined as the maximum of five p-values for the five epitopes, and the dominant epitope is defined as the corresponding epitope. This definition, i.e. assumption, has lead for H3N2 to vaccine effectiveness predictions that correlate with those observed [2].

Another sequence-based antigenic distance measure uses the fraction of mutated amino acid in all the five epitopes

$$p_{\text{all-epitope}} = \frac{\text{number of mutations in all the five epitopes}}{\text{number of amino acids in all the five epitopes}}. \tag{2.9}$$

As an alternative to p_{epitope} and $p_{\text{all-epitope}}$, p_{sequence} is also used with the definition

$$p_{\text{sequence}} = \frac{\text{number of mutations in the HA1 domain of hemagglutinin}}{\text{total number of amino acids in the HA1 domain of hemagglutinin}}. \tag{2.10}$$

2.2.4 Antigenic Distance Measured by Hemagglutination Inhibition

The animal model method to determine the distance between the vaccine strain and the dominant circulating strain employs the HI assay to give the HI table. See Table 2.1: Here H_{ij}, $i, j = 1, 2$ are four HI titers measuring the capability of antibody j to inhibit hemagglutinin i. Note that in reality, health authorities including WHO and CDC provide HI tables with at least eight antisera to evaluate the antigenic distance between candidate vaccine strains and dominant circulating strain. These HI tables are mathematically equivalent to several 2×2 HI tables each of which defines the

Figure 2.1 : HA1 domain of the H1 hemagglutinin in the ribbon format (PDB code: 1RU7). Epitope A (blue), B (red), C (cyan), D (yellow), and E (red) are space filling. These five H1 epitopes are the analogs of the well-defined H3 epitopes [1].

Table 2.1 : HI table with two strains and four HI titers.

	Ferret antisera against Strain 1	Ferret antisera against Strain 2
Strain 1	H_{11}	H_{12}
Strain 2	H_{21}	H_{22}

antigenic distance between one pair of strains in the original HI table. For each pair of strains, we picked up four entries determined by the identities of these two strains and the two corresponding antisera from the original HI table. The 2×2 HI tables in this manuscript are used to elaborate the formulae for d_1 and d_2. In this context Strain 1 is the vaccine strain and Strain 2 is the dominant circulating strain. Two distance measures have been derived from these four HI titers in the HI table [29, 99]:

$$d_1 = \log_2\left(\frac{H_{11}}{H_{21}}\right) \tag{2.11}$$

$$d_2 = \sqrt{\frac{H_{11}H_{22}}{H_{21}H_{12}}}. \tag{2.12}$$

Note that antigenic cartography is carried out on the asymmetrical distance, d_1 [30]. When the vaccine strain and the dominant circulating strain in one season were not identical, we searched the literature for the HI tables with these two strains. The d_1 and d_2 values were averaged if multiple HI tables were found for one season.

Table 2.2 : Summary of results. Nine pairs of vaccine strains and dominant circulating strains in seven flu seasons in the Northern hemisphere were collected from literature. The quantities n_u, N_u, n_v, N_v, p_{epitope}, $p_{\text{all-epitope}}$, p_{sequence}, d_1, and d_2 are defined in Materials and Methods. Only those seasons when H1N1 virus was dominant in at least one country or region where vaccine effectiveness data were available were considered. Two different vaccines have occasionally been adopted in different geographic regions for the same season, in which case two sets of data were added in this table. An asterisk signifies that co-circulating H3N2 was also found in the same country or region in that season; however, the interference to the final result from H3N2 is expected to be small, and so the sets of data with a single asterisk were preserved.

Season	Vaccine strain	Dominant Circulating strain‡	Vaccine effectiveness (%)	n_u	N_u	n_v	N_v	Dominant epitope	p_{epitope}	$p_{\text{all-epitope}}$	p_{sequence}	d_1	d_2
1982–83	A/Brazil/11/78	A/England/333/80	37.0 ± 12.0[1]	48	118	31	121[1]	A	0.083	0.0311	0.0184	0[10]	1.41[10]
1983–84	A/Brazil/11/78	A/Victoria/7/83	38.1 ± 10.3[1–3]	30 55	60 298	21 46	67[1] 300[2]	C	0.121	0.0497	0.0337	1.13[11–13]	13.66[11,13]
1986–87 (a)	A/Taiwan/1/86	A/Taiwan/1/86	64.8 ± 14.3[3,4]	11	217	13	723[4]		0	0	0	0	1
1986–87 (b)	A/Chile/1/83	A/Taiwan/1/86	18.5 ± 12.1[5]	92	878	75	878[5]	B	0.318	0.0807	0.0399	4[12,14–18]	24.48[14,16–18]
1988–89	A/Taiwan/1/86	A/Taiwan/1/86	43.1 ± 10.0[3,5]	119	1125	89	1126[5]		0	0	0	0	1
1995–96 (a)	A/Texas/36/91	A/Texas/36/91	60.0 ± 27.8[6]	6	12	2	10[6]		0	0	0	0	1
1995–96 (b)*	A/Singapore/6/86	A/Texas/36/91	32.2 ± 5.8[7]	99 176	652 652	57 149	684[7] 684[7]	A	0.125	0.0559	0.0307	0.86[14,19,20]	2.43[14,20]
2006–07	A/New Caledonia/20/99	A/New Caledonia/20/99	40.5 ± 2.5[8]	1085	230729	1221	436600[8]		0	0	0	0	1
2007–08*	A/Solomon Islands/3/2006	A/Solomon Islands/3/2006	62.8 ± 12.6[9]	94	262	8	60[9]		0	0	0	0	1

‡Multiple strains are circulating in each season, while each strain has a specific proportion in the virus population in a certain region and season. The strain with the greatest proportion is defined as the dominant circulating strain, which is listed in this table. The dominant circulating strains in this table were chosen based on the literature on vaccine effectiveness, which also gave the region where the effectiveness data were collected.

Literature used in the metaanalysis: 1. [100]; 2. [101]; 3. [102]; 4. [103]; 5. [104]; 6. [105]; 7. [106]; 8. [107]; 9. [108]; 10. [109]; 11. [110]; 12. [29]; 13. [111]; 14. [112]; 15. [113]; 16. [114]; 17. [115]; 18. [116]; 19. [117]; 20. [118].

2.3 Results

We performed a metaanalysis of identities of the vaccine strains and dominant circulating strains, vaccine effectiveness, and antigenic distances between vaccine strains and dominant circulating strains measured with the HI assay using ferret antisera. In one season dominated by H1N1, epidemiological statistics in a certain region reported in literature was used to fix the values of n_u, N_u, n_v, N_v, and the mean and standard error of the vaccine effectiveness. HI assay data in literature are also used to determine antigenic distance d_1 and d_2 between the vaccine strain and dominant circulating strain. Results of the metaanalysis are listed in Table 8.2. Sequence-based antigenic distances p_{epitope}, $p_{\text{all-epitope}}$, and p_{sequence} are calculated from the sequences of the vaccine strain and dominant circulating strain by equations 2.8, 2.9, and 2.10, respectively. Values of p_{epitope}, $p_{\text{all-epitope}}$, and p_{sequence} in each season dominated by H1N1 are also listed in Table 8.2.

While the number of data points is limited, a linear relationship exists between vaccine effectiveness and p_{epitope} by using least squares. Similar to the case for H3N2 influenza [2], p_{epitope} strongly correlates with H1N1 vaccine effectiveness, with $R^2 = 0.68$. The fitted model predicts a vaccine effectiveness of 52.7% when $p_{\text{epitope}} = 0$, and vaccine effectiveness is greater than zero when $p_{\text{epitope}} < 0.442$. In Figure 2.2, the fitted trend line is within one standard error of all data points with $p_{\text{epitope}} > 0$, validating the ability of the p_{epitope} model to predict the vaccine effectiveness with only the sequences of the vaccine strain and the dominant circulating strain.

Although statistical errors exist in the observed vaccine effectiveness, the collected vaccine effectiveness data reject the null hypothesis that the vaccine effectiveness is independent of p_{epitope}. The nine pairs of vaccine strains and dominant circulating strains in Table 8.2 have five difference antigenic distances between vaccine strain and

dominant circulating strain defined by p_{epitope}. The nine pairs of strains were thus categorized into group 1–5 with p_{epitope} equal to 0, 0.083, 0.121, 0.125, and 0.318, respectively, and the average vaccine effectiveness and standard error were calculated for each group. The vaccine effectiveness differences between these five groups were significant, such as group 1 and group 4 ($p = 0.0079$) and group 1 and group 5 ($p = 0.0054$). Moreover, statistical analysis shows that the introduction of p_{epitope} is valuable in the selection process of vaccine strains. The slope of the fit line is significantly smaller than zero ($p = 0.0027$). Hence the linear model is able to predict the vaccine effectiveness with the knowledge of p_{epitope}. In other words the non-zero slope of vaccine effectiveness as a function of p_{epitope} is significant to the level of 0.27%.

Two other sequence-based antigenic distance measures alternative to p_{epitope} are $p_{\text{all-epitope}}$ and p_{sequence}. Unlike p_{epitope}, which focuses upon the mutations in the antibody binding regions, $p_{\text{all-epitope}}$ calculates the fraction of mutated amino acids in all the five epitopes, and p_{sequence} calculates the fraction of mutated amino acids in the whole HA1 domain of hemagglutinin. The p_{sequence} measure is also one of the optional distance measures for phylogenetic softwares. In Figure 2.3, the correlation between H1N1 vaccine effectiveness and $p_{\text{all-epitope}}$ has $R^2 = 0.70$. In Figure 2.4, the correlation between H1N1 vaccine effectiveness and p_{sequence} has $R^2 = 0.66$. The predicted 54% vaccine effectiveness when $p_{\text{all-epitope}}$ in Figure 2.3 and when $p_{\text{sequence}} = 0$ in Figure 2.4 are almost the same as the 53% predicted by the p_{epitope} method. By contrast $p_{\text{all-epitope}}$ and p_{sequence} for H3N2 have less impressive correlations with H3N2 vaccine effectiveness [61, 2], and $p_{\text{all-epitope}}$ and p_{sequence} are not as effective as p_{epitope} as antigenic distance measures and vaccine effectiveness predictors for H3N2.

The HI assay and derived distance measures d_1 and d_2 are still the most widely used measures by researchers and health authorities to identify newly collected cir-

culating strains. These methods are used to recommend the vaccine strain for the coming flu season [119, 120, 121], to draw the antigenic map [30], and to support the phylogenetic data [119]. Figure 2.5 and 2.6 describe the correlation between vaccine effectiveness and antigenic distances d_1 and d_2 from the HI assay. A correlation is found in both figures. In the season 1995–96 in Israel, the vaccine strain is A/Singapore/6/86 (H1N1) and the dominant circulating strain is A/Texas/36/91 (H1N1), between which the averaged d_1 is 0.86. Since the vaccine effectiveness is only 32.2%, its discrepancy to the corresponding effectiveness 42.5% in the trend line is much larger than one standard error of vaccine effectiveness. Similarly, the same pair of vaccine strain and dominant circulating strain introduces a data point further from the trend line if d_2 is used as the distance measure. We also notice that two strains could be antigenically identical as measured with HI assay but antigenically distinct as measured with p_{epitope}. As shown in Table 8.2, in the season 1982–1983, the H1N1 vaccine strain A/Brazil/11/78 and dominant circulating strain A/England/333/80 presented the antigenic distance measured with HI assay $d_1 = 0$ and the sequence-based antigenic distance measure $p_{\text{epitope}} = 0.083$. The H3N2 vaccine strain and dominant circulating strain showed identical d_1 and d_2 values but distinct p_{epitope} values in the seasons 1996–1997 and 2004–2005 [2]. Note that if p_{epitope} is incorporated into the linear models shown in Figure 2.5 and 2.6, the R^2 value is increased. We fit a linear model vaccine effectiveness $= \alpha + \beta_1 p_{\text{epitope}} + \beta_2 d_1 + \beta_3 d_2 + \epsilon$ in which ϵ is an error term. The fitted model is vaccine effectiveness $= 0.54 - 2.179 p_{\text{epitope}} + 0.068 d_1 + 0.003 d_2$ with $R^2 = 0.72$.

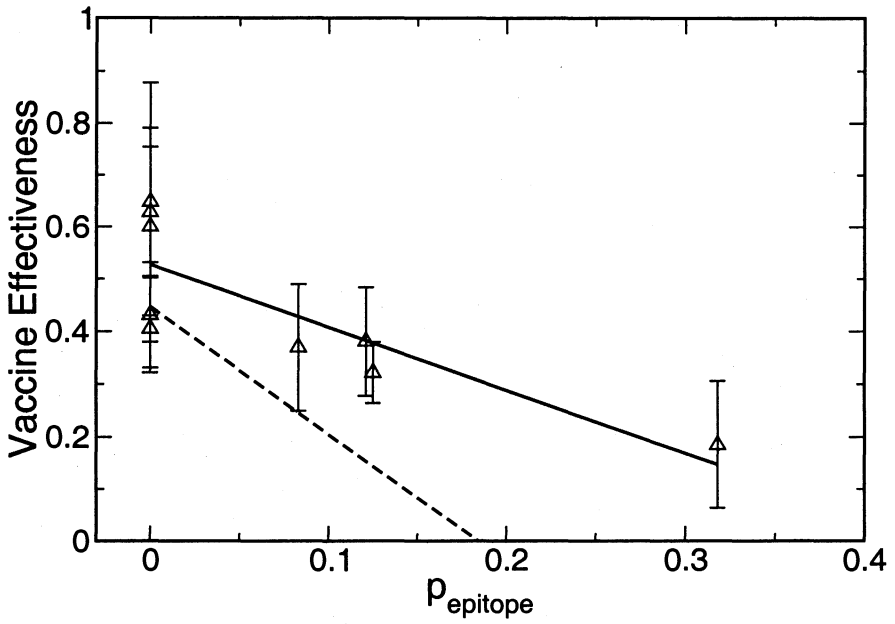


Figure 2.2 : Vaccine effectiveness for influenza-like illness correlates with p_{epitope}, $R^2 = 0.68$ (solid line). Data from Table 8.2. The trend line quantifies vaccine effectiveness as a decreasing linear function of p_{epitope}. Vaccine effectiveness $= -1.19\, p_{\text{epitope}} + 0.53$. Also shown is the vaccine effectiveness to H3N2 (dashed line) [2].

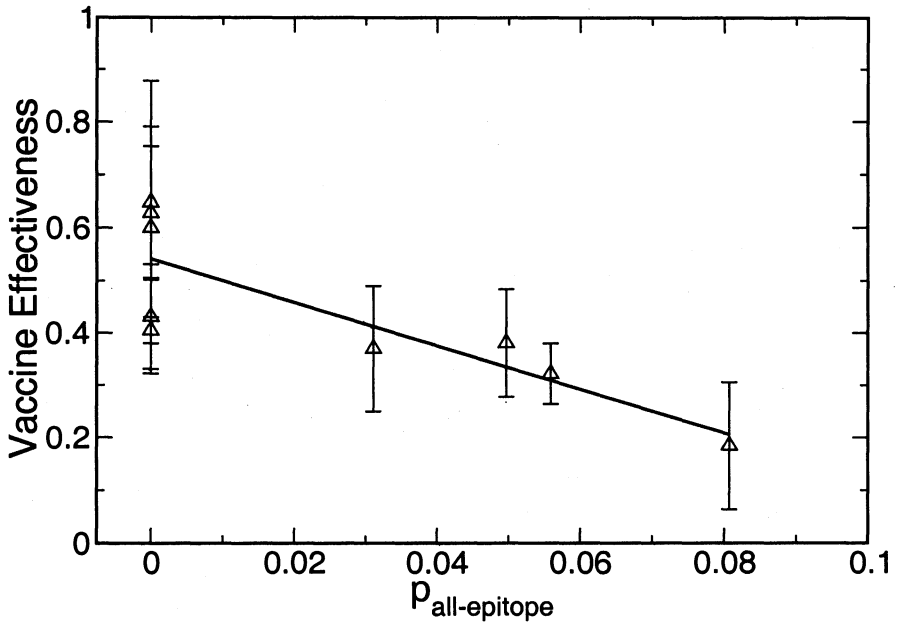


Figure 2.3 : Vaccine effectiveness for influenza-like illness correlates with $p_{\text{all-epitope}}$ with $R^2 = 0.70$. Data from Table 8.2. The trend line quantifies vaccine effectiveness as a decreasing linear function of $p_{\text{all-epitope}}$. Vaccine effectiveness $= -4.16\ p_{\text{all-epitope}} + 0.54$.

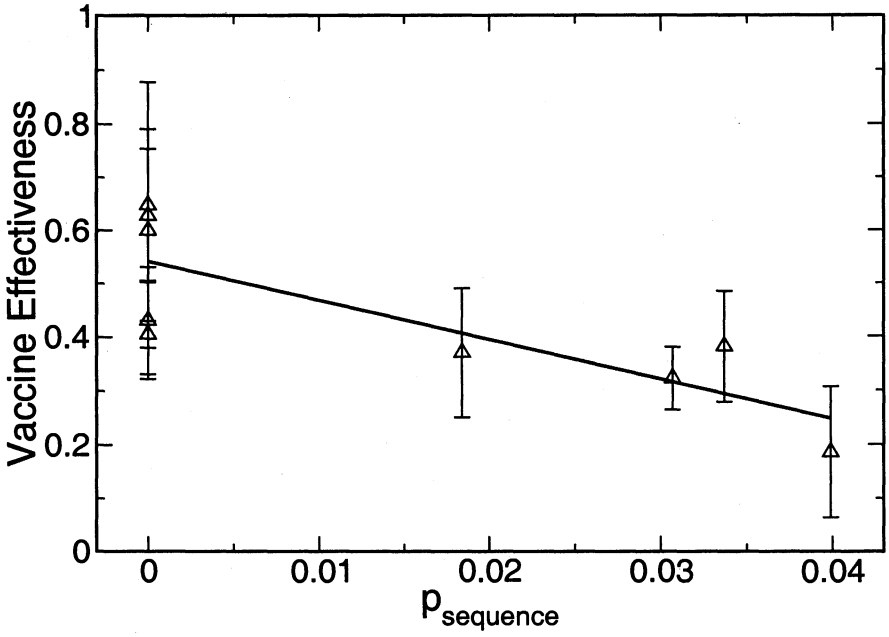


Figure 2.4 : Vaccine effectiveness for influenza-like illness correlates with p_{sequence} with $R^2 = 0.66$. Data from Table 8.2. The trend line quantifies vaccine effectiveness as a decreasing linear function of p_{sequence}. Vaccine effectiveness $= -7.37\ p_{\text{sequence}} + 0.54$.

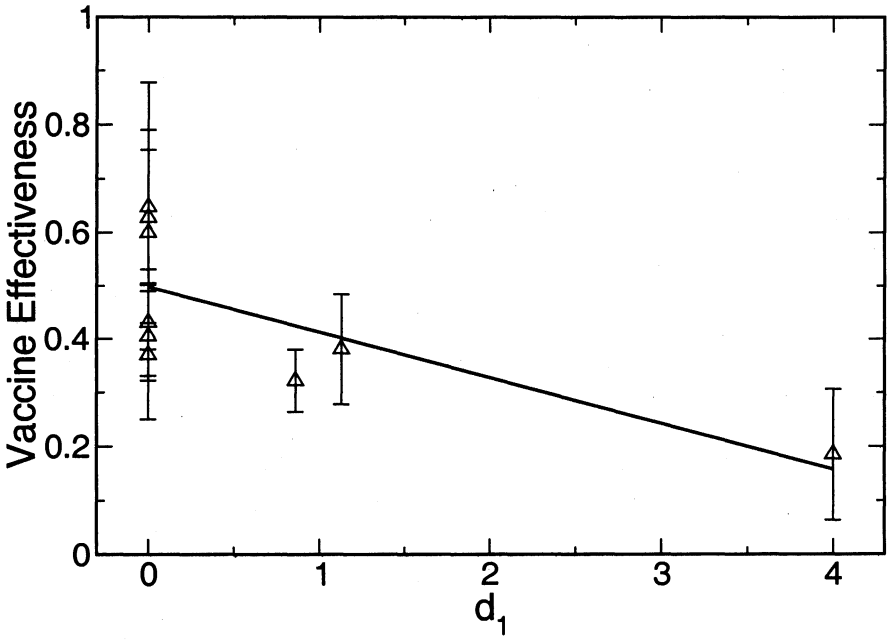


Figure 2.5 : The correlation with $R^2 = 0.53$ between vaccine effectiveness for influenza-like illness and d_1, the antigenic distance defined by HI assay using ferret antisera. Data from Table 8.2. The d_1 values were averaged if multiple HI assay experimental data were found. The trend line quantifies vaccine effectiveness as a decreasing linear function of d_1. Vaccine effectiveness $= -0.085\ d_1 + 0.50$.

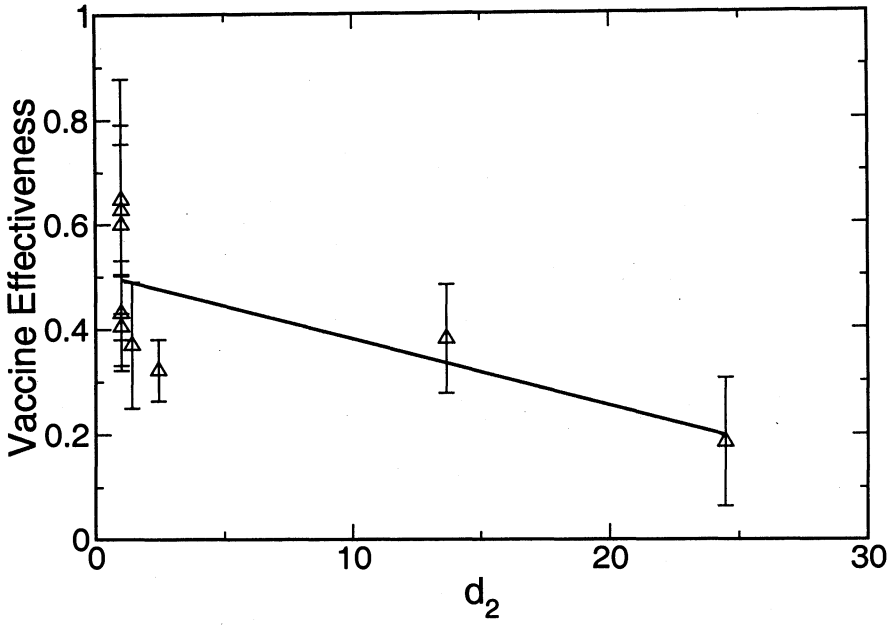


Figure 2.6 : The correlation with $R^2 = 0.46$ between vaccine effectiveness for influenza-like illness and d_2, the antigenic distance defined by HI assay using ferret antisera. Data from Table 8.2. The d_2 values were averaged if multiple HI assay experimental data were found. The trend line quantifies vaccine effectiveness as a decreasing linear function of d_2. Vaccine effectiveness $= -0.013\ d_2 + 0.51$.

2.4 Discussion

2.4.1 Verification of the p_{epitope} Model

Originally the p_{epitope} model was implemented for the H3N2 virus, where p_{epitope} correlates with H3N2 vaccine effectiveness with a significantly larger R^2 than do $p_{\text{all-epitope}}$ and p_{sequence} [2, 61]. In the case of H1N1, the advantage of p_{epitope} over $p_{\text{all-epitope}}$ and p_{sequence} is not as remarkable as for H3N2. We speculate that antibodies against the H3N2 virus may bind to a small fixed region on the surface of H3 hemagglutinin while antibodies against the H1N1 virus may have multiple binding regions available. In other words, we speculate that the dominant epitope in H3 hemagglutinin may contribute substantially to the escape of the H3N2 virus from host antibodies, while escape mutations may occur in the dominant epitope as well as perhaps the subdominant epitopes of H1 hemagglutinin. Our speculation comes from the fact that the epitope region in H1N1 contains more amino acid positions than does that in H3N2 [1].

Two recent epidemiological studies [122, 123] present further support of the p_{epitope} model. Before the emergence of the H1N1 pandemic flu in April 2009, the 2008–2009 flu season was dominated by subtype H1N1 seasonal flu. Both the dominant circulating strain and the vaccine strain in the 2008–2009 season were A/Brisbane/57/2007 (H1N1) [124]. The observed vaccine effectiveness against seasonal flu was 44% (95% CI: 33% to 59%) [122]. The p_{epitope} model predicts the vaccine effectiveness as 53%, which falls into the 95% CI of the reported vaccine effectiveness.

After April 2009, a new peak of influenza activity emerged. The dominant circulating strain in this period was the pandemic H1N1 strain A/California/7/2009 [125, 126]. The reported effectiveness of the 2008–2009 seasonal flu vaccine against

the H1N1 pandemic flu was -50% to -150% [122] and -10% (95% CI: -43% to 15%) [123]. The value of p_{epitope} between A/California/7/2009 and A/Brisbane/57/2007 is 0.77 with epitope B as the dominant epitope. The vaccine effectiveness forecast by the p_{epitope} model is -39%, which agrees with the measured vaccine effectiveness values.

2.4.2 Comparison of H3N2 and H1N1 Vaccine Effectiveness and Evolution Rates

The p_{epitope} model has been previously applied to the prediction of H3N2 vaccine effectiveness [2]. The H3N2 vaccine effectiveness with $p_{\text{epitope}} = 0$ is 44.6%, and vaccine effectiveness is greater than zero for $p_{\text{epitope}} < 0.184$ [2]. Thus, H1N1 vaccines tend to have higher vaccine effectiveness compared to H3N2 vaccines, as shown in Figure 2.2. The comparison between H3N2 and H1N1 vaccine effectiveness (Figure 2.2 versus Figure 2 of [2]) illustrates that H1N1 vaccine has higher effectiveness than the H3N2 vaccine as a function of p_{epitope}. This observation suggests that the host immune system is more effective at recognizing and eliminating the H1N1 virus ($p_{\text{epitope}} = 0$), and that humoral cross immunity is stronger for H1 hemagglutinin ($p_{\text{epitope}} > 0$). This observation also explains why an H3N2 epidemic is usually a more severe health threat than an H1N1 epidemic. We propose that H1N1 has a longer history of circulating in the human population, so human immune system may recognize H1N1 more effectively, and this may be the reason that under stronger immune pressure, the H1N1 virus may have a higher degree of adaptation to the human host. In the following discussion, we verify this hypothesis by two facts. First, the H1N1 virus has a larger antigenic diversity than does the H3N2 virus. Second, the H1N1 virus presents higher evolutionary rate in the per dominant season basis.

To compare the antigenic diversities of H1N1 and H3N2, we downloaded from the

NCBI database on 13 August 2009 all the amino acid sequences of H3 hemagglutinin collected in the 18 years with H3N2 dominant circulating strains [2] and those of H1 hemagglutinin collected in 7 years with H1N1 dominant circulating strains (Table 8.2). Thus 18 subsets of H3N2 sequences and 7 subsets of H1N1 sequences were formed. The centers of these subsets are the corresponding vaccine strains in the same season of the circulating virus. The radius of each subset is obtained by the calculation of p_{epitope}. First, the strains with the top 5% p_{epitope} antigenic distance measure to the center of each subset were selected, to focus on the extent of viral evolution. Second, the p_{epitope} between these selected strains and the center were averaged in each year as the radius. Third, the radii were averaged over all the 18 years for H3N2 and over 7 years for H1N1. That is, the average radius of the top 5% was calculated in each year. As a result, the average H3N2 subset radius with the vaccine strains as the centers is 0.211. The average H1N1 radius is 0.520 with the vaccine strains as the centers. This difference between the H3N2 radius and the H1N1 radius is significant with the p-value 0.0118 using the Wilcoxon rank-sum test. Consequently, the H1N1 virus has a larger antigenic diversity in each season compared to the H3N2 virus, as shown in Figure 2.7.

We also compared the evolutionary rates of H1N1 and H3N2 because evolutionary rate of the virus is an index of the selection pressure of the virus. The virus undergoes less immune pressure in a non-dominant season and high immune pressure in a dominant season. It has been noticed that in H1 and H3 hemagglutinin, the region outside epitopes presents significantly lower evolutionary rate than do the epitopes [1, 42]. This phenomenon indicates that without immune pressure, the spontaneous evolutionary rates of both H1N1 and H3N2 are low. Therefore, a higher evolutionary rate of one virus subtype in a dominant season comes from the higher immune

pressure rather than neutral evolution, and we reject the alternative scenario that the higher evolutionary rate causes a virus subtype to be dominant in one season. So the evolutionary rate per dominant season is a natural measure of the virus evolution. Between 1983 to 1997, H3N2 was dominant in 8 of 15 years, and between 1977 to 2000, H1N1 was dominant in 5 of 24 years [42]. Between 1980 to 2000, the HA1 domain of H3 hemagglutinin has a higher annual evolutionary rate of 3.7×10^{-3} nucleotide substitution/site/year than does the HA1 domain of H1 hemagglutinin, which has the annual evolutionary rate of 1.8×10^{-3} nucleotide substitution/site/year [42]. Measured on a per dominant season basis, however, the HA1 domain of H1 hemagglutinin evolves faster in its dominant season with the rate of 8.6×10^{-3} nucleotide substitution/site/dominant season than does the H3 hemagglutinin with the rate of 6.9×10^{-3} nucleotide substitution/site/dominant season. The difference is significant with a p-value 0.0008. Similarly, between 2000 and 2007, the HA1 domain of H1 hemagglutinin evolves faster in its dominant season with the rate of 10.2×10^{-3} nucleotide substitution/site/dominant season than does the H3 hemagglutinin with the rate of 7.4×10^{-3} nucleotide substitution/site/dominant season. The difference is significant with a p-value 0.0005 [127]. Here we have divided the annual evolutionary rate by the proportion of dominant years for both H1 and H3 hemagglutinin. Even on a short time scale without fixation, H1 hemagglutinin shows a comparable or higher mutation rate of 9.1×10^{-6} nucleotide substitution/site/day than H3 hemagglutinin of 4.2×10^{-6} nucleotide substitution/site/day ($p = 0.26$) [37], probably caused by the adaptation to the higher immune pressure, at least for some strains. To make this last point, we have assumed that the mutation rate of the HA gene is the same as that of the NS gene. We assume that the same polymerase is operating on these two genes, and so the mutation rates are expected to be the same. The comparisons of

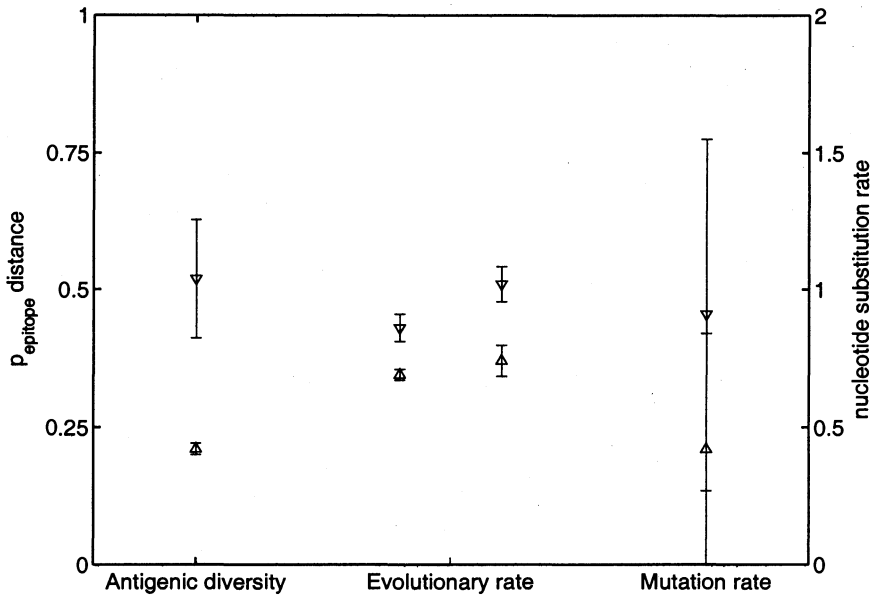


Figure 2.7 : The comparison between H3N2 (triangle up) and H1N1 (triangle down) in regard to the antigenic diversity, the evolutionary rate between 1980 and 2000 (left), the evolutionary rate between 2000 to 2007 (right), and the mutation rate on a short time scale without fixation. The antigenic diversity is measured with p_{epitope}, the unit of evolutionary rate is 10^{-3} nucleotide substitution/site/year, and the unit of mutation rate is 10^{-6} nucleotide substitution/site/day.

evolutionary rates and mutation rates between H3N2 and H1N1 are summarized in Figure 2.7.

2.4.3 The p_{epitope} Model as a Supplement to HI Assay

For both H1N1 (this chapter) and H3N2 [2], the HI assay correlates less well with vaccine effectiveness than does p_{epitope}. Collection of HI assay data measuring antigenic distance is also more time-consuming and more expensive compared to the p_{epitope} model. Many hundreds of strains are circulating and collected in an average flu season, thus an HI table with tens of thousands of entries needs to be built to assess the

antigenic distance between each pair of strains. With the high-throughput sequencing technology generating hemagglutinin sequence data, such antigenic distances are easily measured with the sequence-based antigenic distance measure p_{epitope}, which correlates to a greater degree with vaccine effectiveness than do the HI data.

The p_{epitope} model is developed to provide researcher and health authorities with a new tool to quantify antigenic distance and design the vaccine. We do not suggest that p_{epitope} should substitute for the current HI assay, but rather suggest that p_{epitope} serves as an additional assessment when selecting vaccine strains. Using p_{epitope} to supplement to HI assay data may allow researchers and health authorities to more precisely quantify the antigenic distance between dominant circulating strains and candidate vaccine strains. The adoption of the p_{epitope} theory may also allow researchers to minimize the cost and the number of ferret experiments and to correct HI assay data in some situations.

2.5 Supplementary Materials

2.5.1 Humoral Immune System Plays a Major Role in Immunity to Influenza

The influenza vaccine considered in this study is the trivalent inactivated vaccine (TIV) administered by intramuscular injection. The effective components of TIV are hemagglutinin (HA) and neuraminidase (NA) that noticeably induce the humoral immunity but activate the cellular immunity less vigorously [128]. The other, cold–adaptive trivalent live attenuated influenza vaccine (LAIV) is also believed to induce the cellular immunity to a low level [128].

The humoral immunity greatly relies on the antigenic distance between the hemag-

glutinin of the vaccine and that of the dominant circulating strain. On the other hand, the cellular immune system focuses on the highly conserved internal proteins, which are the Matrix protein 1 (M1) and the nucleoprotein (NP) [73]. In contrast to the antibodies, CD8+ and CD4+ T cells show notable cross immunity to a wide variety of strains [73]. Like the cellular immune system, the antigen-unspecific innate immune system generates a homogeneous immune reaction against different influenza strains [72].

For all these reasons ferret antisera, in which antibodies are the major immune component, is used in the hemagglutination inhibition (HI) assay as the conventional way to measure the antigenic distance between the vaccine strain and the dominant circulating strain. Therefore, we consider the antibody rather than the cellular or innate immune system to be the dominant element in our quantification of antigenic distance between two influenza strains and the key factor for influenza vaccine effectiveness.

2.5.2 Evaluation of Vaccine Effectiveness

By definition, vaccine efficacy is measured by controlled trials with initially susceptible subjects, while vaccine effectiveness is measured by epidemiological observance of susceptible population without giving placebo [129]. Vaccine efficacy is relatively more idealized than vaccine effectiveness, because vaccine effectiveness depends on vaccine efficacy and other environmental factors [130]. Although the terms vaccine efficacy and vaccine effectiveness are interchangeable to some extent [130], we use the term vaccine effectiveness because factors other than vaccine strain and dominant circulating strain are involved in the studies used in our metaanalysis. The data source for vaccine effectiveness calculation used ILI as the primary endpoint, and

studies we use contained controlled unvaccinated groups.

The data from four studies require additional clarification. The paper by Edwards et al. [104] did not provide the retrospectively reported influenza–like illness data prior to the 1987–88 season, so we use the number of ill subjects presenting for throat culture to calculate the morbidity rate u and v. Note that in this study, subjects with influenza–like disease were required to show up for a throat culture, and when characterized by vaccination status the numbers of such patients is thus a reasonable estimation of the illness rate u and v. Moreover, in other seasons when both retrospective data and number of presenting ill subjects were available, the vaccine effectiveness calculated from retrospective data and numbers of presenting ill subjects are similar to each other, especially for subtype H1N1 [104]. In Grotto et al.'s study [106] using influenza strains from Israel in the 1995–96 season, the numbers of sampled H1N1 and H3N2 strains in Israel was 7 and 35, respectively. The samples were collected from six clinics in December that was in the middle of the influenza season. However, the number of both subjects and viruses sampled are limited. At the global level with more available data, it was observed that H1N1 and H3N2 were co-circulating with comparable frequencies, and H1N1 virus was found in North America and part of Eurasia including Israel [131]. The proportion of H1N1 in samples during the 1995–96 season in USA ranks #5 in H1N1 proportion since 1977 [42]. In the same season, H3N2 vaccine strain and dominant circulating strain were a perfect match, so the decrease in the overall vaccine effectiveness is expected to be due to the mismatch in the H1N1 component. Therefore we treat H1N1 here as a co-circulating strain and take into account the vaccine effectiveness reported in this article [106]. Keitel et al. [103] reported that the dominant circulating strain in the 1983–84 season was A/Chile/1/83 rather than A/Victoria/7/83 in this table and in

other cited studies. The illness rate u and v are small, and so the standard error of vaccine effectiveness is 64.8%, which is unacceptable. Thus the use of Keitel et al.'s data for 1983–84 season is not appropriate to the vaccine effectiveness assessment. The reference by Couch et al. [102] did not provide original data n_u, N_u, n_v, and N_v for the calculation of vaccine effectiveness. Error bars of vaccine effectiveness in these seasons were calculated with other data sources.

2.5.3 Robustness of the p_{epitope} Model

Influenza vaccine effectiveness may depend not only on the antigenic distance between the vaccine strain and the dominant circulating strain quantified by p_{epitope}, but also on the percentage of people vaccinated, the time of vaccination in the influenza season, influenza virus transmissibility and reproduction rate, and individual's immune history. Thus, development of the public health system and a greater fraction of the population being vaccinated may result in a trend of both H1N1 and H3N2 vaccine effectiveness. The statistics of vaccine effectiveness could be biased by these factors. Nevertheless, greater than 50% of the H1N1 and H3N2 vaccine effectiveness are explained by p_{epitope}, since $R^2 > 1/2$. To show that the model of vaccine effectiveness can be well reduced to a linear form between p_{epitope} and vaccine effectiveness, we calculated the residuals of linear regression of vaccine effectiveness on p_{epitope}, and performed another linear regression of these residuals versus year. The trend line of the residuals has a slope of -0.0002/year and the null hypothesis that the slope equals zero is not rejected ($p = 0.96$), as shown in Figure 2.8. The residuals of H3N2 vaccine effectiveness [2] were also correlated with the year and the slope -0.0013/year is not significantly different with zero ($p = 0.58$), as shown in Figure 2.9. Therefore the contribution of other simple time–dependent factors other than p_{epitope} to H1N1

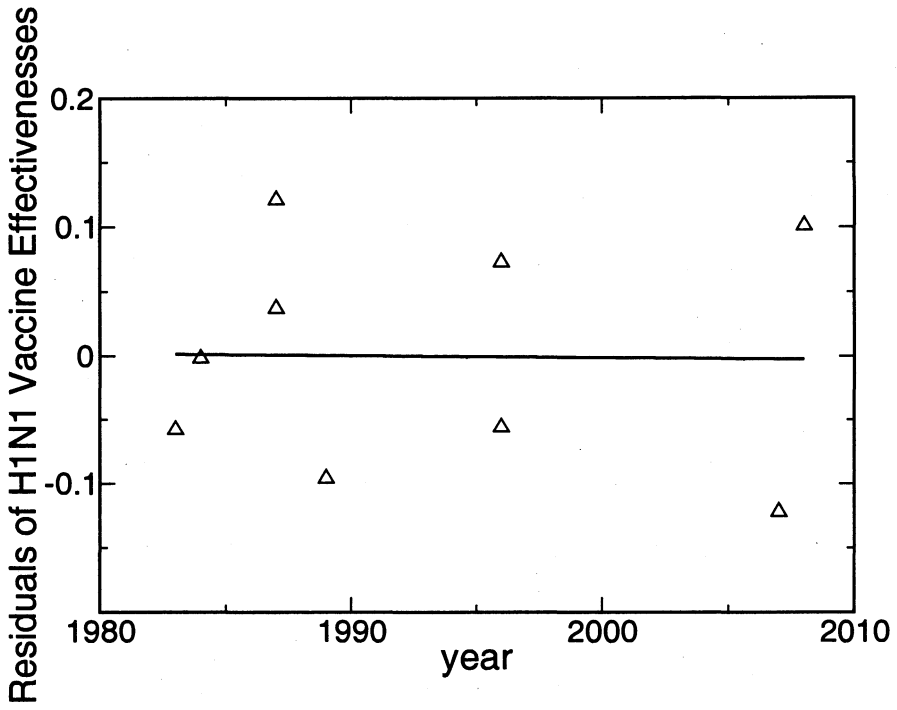


Figure 2.8 : The linear regression with $R^2 = 0.0003$ of the residuals of H1N1 vaccine effectiveness versus year. Data from Table 1 and Figure 2 in the main text. The slope of the trend line is -0.0002/year. ANOVA test: H_0: slope $= 0$, $F = 0.0021$, and $p = 0.96$. The null hypothesis that these residuals are independent of time cannot be rejected.

and H3N2 vaccine effectiveness in humans is negligible. Our analysis suggests that the vaccine effectiveness data in this chapter are negligibly affected by these potential biases.

Despite the limited number of available data points in this study, the correlation line between the p_{epitope} and the vaccine effectiveness has statistical meaning. In Figure 2 in the main text, the trend line is vaccine effectiveness $= -1.19\, p_{\text{epitope}} + 0.53$, the greatest determinant of which is the data point 1986–87 (b). If this data point is removed, the trend line becomes vaccine effectiveness $= -1.63\, p_{\text{epitope}} + 0.54$, which is

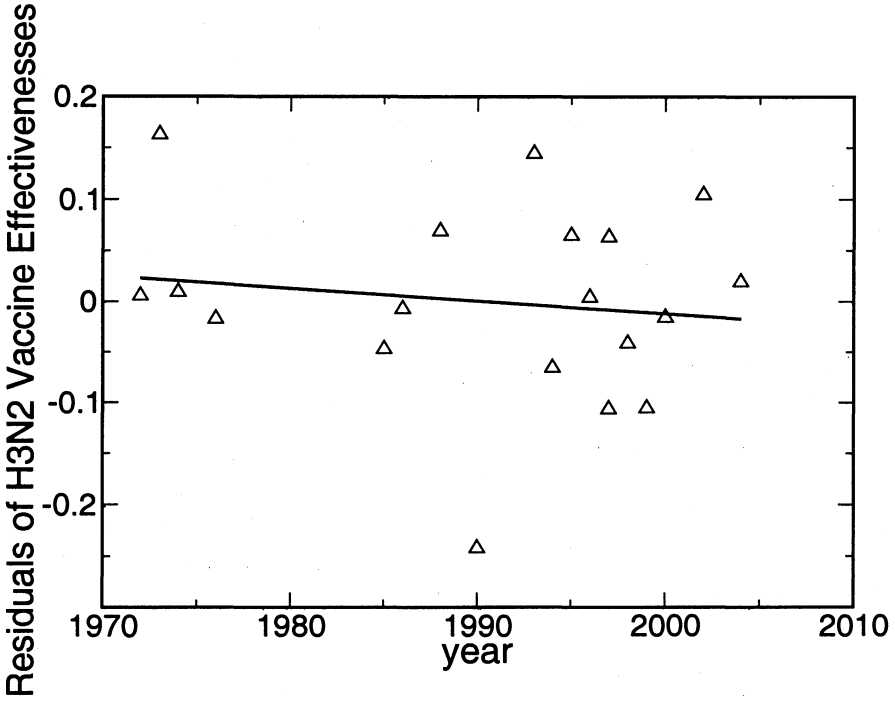


Figure 2.9 : The linear regression with $R^2 = 0.018$ of the residuals of H3N2 vaccine effectiveness versus year. Data from [2]. The slope of the trend line is -0.0013/year. ANOVA test: H_0: slope $= 0$, $F = 0.32$, and $p = 0.58$. The null hypothesis that these residuals are independent of time cannot be rejected.

not fundamentally distinct with the original trend line. In the data point 1986–87 (b), the difference of the vaccine effectiveness predicted by these two models is 0.13, which is roughly one standard error. In reality, most p_{epitope} values are less than 0.1, and so most of the differences between these two predicted vaccine effectiveness values are less than 0.034, which is within the noise levels of the epidemiological measurements.

The basis for calculating p_{epitope} is a set of well defined epitopes. An early definition of the five epitopes in H1 hemagglutinin [3] did not identify numerous amino acid positions in which mutations were frequently selected in history. These positions are presumably under strong antigenic pressure to be selected for escape mutation. The more recent definition of H1 epitopes incorporates these additional amino acid positions as well as amino acids from the epitopes of H3 hemagglutinin [1]. Likely additional experiments on the H1 epitopes will allow further refinement of the calculation of p_{epitope}. Only nine epidemiological data points are available since the reemergence of H1N1 virus in humans in 1977. The p_{epitope} model parameters may be further improved as epidemiological data are accumulated.

The antigenic properties are determined by a small number of amino acid substitutions, because the positions and the amino acids introduced by mutation have distinct effects on the change of antigenic distance between vaccine strain and dominant circulating strain. For example, mutations yielding charged amino acids in the dominant epitope are favorable for the virus, and may be the key amino acid substitution for the antigenic properties [132]. An improved sequence–based model might assign different amino acid substitutions with weights determined by the decrease of binding constant between HA and antibody using free energy calculation [94]. With the current knowledge, the less precise but safe p_{epitope} model assigns the amino acid substitutions in the dominant epitope with weight one, and assigns other amino acid

substitutions with weight zero. The p_{epitope} model can nevertheless correlate with the vaccine effectiveness better than the antisera data. That is, for both H1N1 and H3N2, the p_{epitope} method is superior to other methods in current use.

2.5.4 Comparison of the p_{epitope} Model and the HI Assay for H3N2 Virus

In some cases p_{epitope} model detects antigenic variants better than the HI assay. In the 2003–04 Northern hemisphere flu season, the majority of isolated H3N2 strains were similar to A/Fujian/411/2002 using HI assay [133], hence WHO recommended a A/Fujian/411/2002-like strain as the 2004–05 Northern hemisphere H3N2 vaccine component, and A/Wyoming/3/2003 was selected. Although A/Wyoming/3/2003 is similar to A/Fujian/411/2002 circulating in 2004–05 ("antigenically equivalent" by HI data [134]), the vaccine effectiveness was only moderate [2]. Interestingly, the p_{epitope} between A/Fujian/411/2002 and A/Wyoming/3/2003 was also moderate ($p_{\text{epitope}} = 0.095$), predicting the moderate vaccine effectiveness. In fact, the p_{epitope} method can also detect antigenic variants more rapidly as they emerge [135].

Chapter 3

Comment on Ndifon et al, "On the use of hemagglutination-inhibition for influenza surveillance: Surveillance data are predictive of influenza vaccine effectiveness"

In 2006, Gupta et al. published an analysis of vaccine efficacy in humans for H3N2 influenza A. We collected vaccine efficacy data from the epidemiological literature for years between 1971 and 2003. In total, 19 efficacy values were obtained. We determined correlations between vaccine efficacy and four measures of antigenic distance. The first measure is the fraction of amino acid differences between the vaccine strain and the dominant circulating strain in the hemagglutinin HA1 sequence, p_{sequence}. The second measure is the fraction of amino acid differences between the dominant epitope of the vaccine and dominant circulating strain, p_{epitope}. The third measure is the logarithm base 2 of the ratio of homologous to heterologous titers, d_1 [29]. The fourth measure is the square root of the ratio of the homologous titers to the heterologous titers, d_2 [99]. The correlations of the different measures of antigenic distance with vaccine efficacy are shown in Table 3.1.

In 2009, Ndifon et al. published an analysis of a subset of these data, 11 data points, with some modifications [136]. Four new data points were added: some early data from 1968 and 1969, data for 1980/1981, and recent data for 2004/2005. Data for 1971/1972, provided in [2], were omitted from [136].

There are a number of discrepancies in the data of Table 1 of Ndifon et al. [136]. In

1972/1973 the vaccine strain was listed as A/Hong Kong/1/68. The correct vaccine strain was A/Aichi/2/68 (also known as X31) [2, 137, 138]. A value of p_{epitope} of 0.263 was listed in [136], rather than the correct 0.190 [2]. In 1994/1995, a dominant circulating strain of A/Shangdong/9/93 was listed [136], rather than a mixture of strains that more closely resemble A/Johannesburg/33/94 than A/Shangdong/9/93 [2, 139], and which we represent by the former [2]. A value of p_{epitope} of 0 was listed in [136], rather than 0.105 [2]. In 1995/1996 the vaccine efficacy was listed as 42.0% [136], rather than 45% [2]. In 1996, vaccine and circulating strains of A/Wuhan/359/95 and A/Nanchang/933/95 were listed [136] instead of the correct A/Nanchang/933/95 vaccine and A/Wuhan/359/95 US CDC-determined circulating strain [2, 140, 141]. In 1997, the dominant circulating strain of A/Nanchang/933/95 was listed [136] instead of A/Wuhan/359/95 [2, 141], leading to p_{epitope} of 0 [136] instead of 0.095 [2]. In 1997/1998 a p_{epitope} value of 0.227 [136] was listed instead of the correct 0.238 [2]. The vaccine discrepancies in [136] stem from the incorrect assumption that the WHO "recommended" strain was administered, rather than the "like" vaccine strain that was actually manufactured and administered. Finally, in 2003/2004 efficacy data for individuals vaccinated within 2 weeks of illness were removed from the dataset, even though it is known that the immune response to influenza reaches quite high levels and influenza virus is suppressed within 2-3 days of exposure. A low efficacy value of 0.8% [136] rather than 12% [2] was reported.

Once these amendments are made, the 23 data in aggregate from [2, 136] reveal a correlation between vaccine efficacy and the p_{epitope} theory of $R^2 = 0.76$ (see Figure 3.1). We focus here on the difference between p_{epitope} and the rAHM measure of antigenic distance reported as correlating well with vaccine efficacy in [136]. We note that the definition of rAHM is identical to that of d_2. In [136], only half of the data

were used, those for which the vaccine and dominant circulating strain were distinct, a total of 11 data points. These 11 data points were used to test the p_{epitope}, p_{sequence}, and d_2 = rAHM measures of antigenic distance. With the corrections discussed above made, there are 14 data points fitting this criterion. If a large amount of vaccine efficacy data were available, removing a small subset of data would not be problematic. Removing 50% of the data, so that there are no data for small to moderate antigenic distances, lead to a number of artifacts. The first artifact is that the linear fit of rAHM to the 11 data points of vaccine efficacy extrapolates to a vaccine efficacy of 18% when the vaccine is identical to the dominant circulating strain, with $R^2 = 0.56$ (see Figure 3.1, left insert). While the R^2 is sizable, the prediction of 18% is discrepant from the average vaccine efficacy of 43% when the vaccine is identical to the dominant circulating strain. That rAHM does not predict moderate antigenic distances well is made clear when the rAHM data are fit to all years, with $R^2 = 0.54$ instead of $R^2 = 0.76$ for the p_{epitope} theory. When p_{epitope} is fit to the amended 14 data points, the linear fit extrapolates to a vaccine efficacy of 27% for an identical vaccine and dominant circulating strain (see Figure 3.1, right insert), with $R^2 = 0.27$, not the nearly zero value reported in [136]. The prediction of the p_{epitope} theory is more accurate than that of the rAHM measure on these out-of-sample data, although the correlation coefficient is lower. We note that 4 out of the 6 points with $p_{\text{epitope}} > 0.19$ have a negative efficacy. This predictive ability is rather similar to that of the rAHM data, for which 4 out of the 5 points with rAHM > 5 have negative vaccine efficacy [136]. In the 2004/2005 season, both A/California/7/2004 and A/Fujian/411/2002 were circulating strains, in addition to a substantial amount of circulating influenza B. The antigenic distance between A/Wyoming/3/2003 and A/California/7/2004 is $p_{\text{epitope}} = 0.286$, and the antigenic

distance between A/Wyoming/3/2003 and A/Fujian/411/2002 is $p_{\text{epitope}} = 0.095$. Thus, while the predicted efficacy for the former is not positive, for the latter it is 20%. Antigenic distance for A/California/7/2004 alone cannot predict the expected vaccine efficacy against multiple nearly-dominant circulating strains. Indeed, the reported efficacy of 9.2% [136] is roughly the average of the 0% and 20% predicted efficacies from the p_{epitope} theory. When the 2004/2005 data point is eliminated, the p_{epitope} prediction extrapolates to 37% efficacy for identical vaccine and dominant circulating strain, with $R^2 = 0.46$.

In summary, the p_{epitope} theory is more accurate and has a larger R^2 value than the rAHM ferret animal model data when all the human H3N2 influenza A vaccine efficacy data are considered. When trained on half of the data, the p_{epitope} theory more accurately predicts the out-of-sample, small and moderate antigenic distance efficacies than does the rAHM data. When the data point for the 2004/2005 season with multiple nearly-dominant circulating strains is removed, the p_{epitope} theory and rAHM data fit have similar R^2 values on years for which the vaccine and dominant circulating strains are distinct. Both p_{epitope} and rAHM predict that vaccine efficacy decreases to zero beyond a critical antigenic distance, given by $p^*_{\text{epitope}} = 0.19$. We note that the p_{epitope} theory requires only sequence information, whereas rAHM is constructed from hemagglutinin inhibition data measured in ferrets.

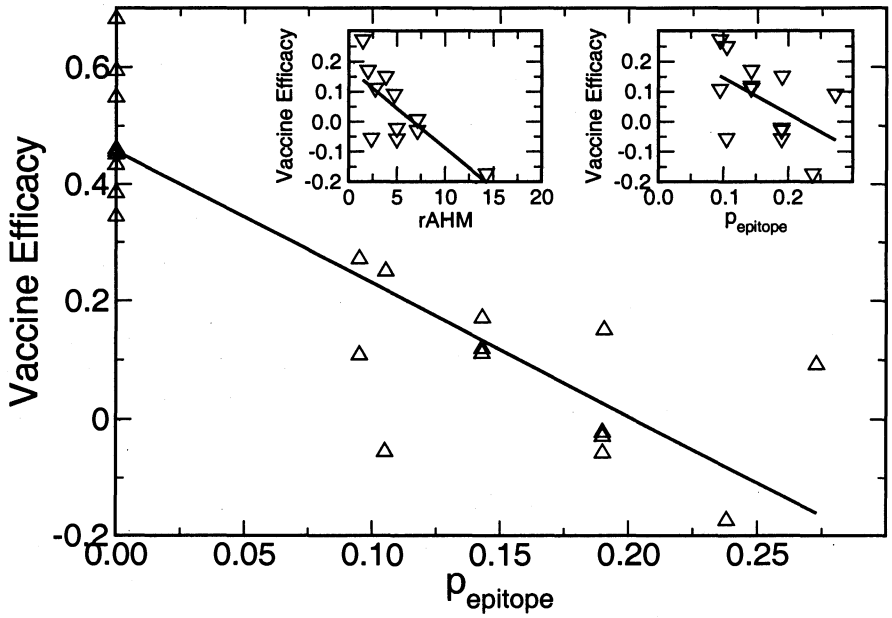


Figure 3.1 : Vaccine efficacy versus the p_{epitope} or rAHM measures of antigenic distance. In inset are the data for which the vaccine and dominant circulating strain are distinct.

Table 3.1 : Correlation of H3N2 Influenza A vaccine efficacy in humans with different measures of antigenic distance.

Method	Correlation	Reference for data
p_{epitope}	0.76	[2, 136]
p_{sequence}	0.59	[2]
d_1	0.57	[2]
d_2	0.43	[2]
rAHM	0.54	[136]

Chapter 4

The Epitope Regions of H1-Subtype Influenza A, with Application to Vaccine Efficacy

The recent emergence of H1N1 (swine flu) illustrates the ability of the influenza virus to create antigens new to the human immune system, even within a given hemagglutinin and neuraminidase subtype. This new H1N1 strain is sufficiently distinct, for example, from the A/Brisbane/59/2007 (H1N1)-like virus strain of influenza in the 2008/09 Northern hemisphere vaccine that protection is not expected to be substantial. The human immune system responds primarily to the five epitope regions of the hemagglutinin protein. By determining the fraction of amino acids that differ between a vaccine strain and a viral challenge strain in the dominant epitope regions, a measure of antigenic distance that correlates with epidemiological studies of H3 influenza A vaccine efficacy in humans with $R^2 = 0.81$ is derived. This measure of antigenic distance is called p_{epitope}. The relation between vaccine efficacy and p_{epitope} is given by $E = 0.472.47 \times p_{\text{epitope}}$. We here identify the epitope regions of H1 hemagglutinin, so that vaccine efficacy may be reliably estimated for H1N1 influenza A.

4.1 Introduction

The recent outbreak of H1N1 (swine flu) has caused immediate international concern. From its earliest case in mid-March 2009 to mid-May 2009, 80009000 infections and 7080 deaths were recorded in 4050 countries and regions, and as of mid-May, over 90% of infections and deaths were in Mexico and the USA [14]. Historically, three

subtypes of influenza A virus have been able to circulate in the human population. The Spanish flu pandemic in 1918–20 was H1N1, which circulated in the world until 1957. H1N1 reappeared in 1977 and persists today [16]. The Asian flu pandemic in 195658 was H2N2, which spread widely in the human population during the time interval 1957–68 [17]. The Hong Kong flu pandemic in 1968–69 was H3N2, which has circulated in the human population as the dominant subtype until recently [17]. Other subtypes rarely infected humans, although cases of H5N1 and H9N2 have been reported.

The 2009 swine flu virus possesses H1 hemagglutinin (HA) and N1 neuraminidase on the surface of the virion, of which the hemagglutinin is the main target of host antibodies. The human immune system responds primarily to the five epitope regions of the hemagglutinin protein [27, 28]. Host antibodies bind to five epitopes in hemagglutinin and lead to high escape evolution rates of amino acids in the epitopes. An early identification of H1 epitopes was carried out by antibody mapping of the A/PR/8/1934 (H1) hemagglutinin, with an additional study of laboratory mutations [3]. However, these H1 epitopes contain far fewer amino acids than do the epitopes in H3 determined by modern methods [28] and are incomplete. Alignment of H1 strains in 1918–2009 indicates many mutation positions outside the originally identified epitopes. We here use sequence alignment and information entropy to complete the definition of H1 epitopes.

Vaccination is an effective way to reduce the influenza morbidity and mortality. The efficacies of influenza vaccines vary from year to year, in part due to different antigenic distances between the circulating influenza strains and the vaccine. Antigenic distance between a vaccine strain and a viral strain can be estimated by the number of mutations in the hemagglutinin sequence between the two strains [29, 30].

Ferret animal model studies are used to further refine the notion of antigenic distance. These methods correlate with epidemiological studies of vaccine efficacy in humans with $R^2 = 0.59$ and 0.43–0.57, respectively [2, 31]. By considering only those mutations that occur in the dominant epitope, the p_{epitope} theory provides a prediction of vaccine efficacy that correlates with epidemiological studies of vaccine efficacy in humans with $R^2 = 0.81$.

Vaccine efficacy has a linear correlation with the antigenic distance between the vaccine strain and the circulating virus strain [2, 31]. Since p_{epitope} correlates well with influenza vaccine efficacy in humans, it can be used to estimate antigenic distance. For example, when p_{epitope} is larger than 0.19, the vaccine no longer offers protection. This correlation can be used to find optimal strains for vaccine design with minimal antigenic distance from expected circulating strains. Here we calculate the antigenic distance for H1 influenza A and apply the method to evaluate efficacy of a candidate swine flu vaccine.

4.2 Methods

4.2.1 Mapping the Epitope from H3 Hemagglutinin to H1 Hemagglutinin

In this chapter, the amino acid positions of H1 (A/PR/8/1934) and H3 (A/Aichi/2/1968) hemagglutinin are denoted using H1 and H3 numbering, respectively [142, 143]. Two hemagglutinin sequences of A/PR/8/1934 (H1) and A/Aichi/2/1968 (H3) are aligned using ClustalW. Amino acids in H1 sequence corresponding to epitope A–E in H3 are defined as mapped epitope A–E in H1. Similarity between H3 epitopes and corresponding mapped H1 epitopes was verified by aligning their three-dimensional structures (PDB code: H3 = 1HGF and H1 = 1RU7). The RMSD values of alpha carbons

in five epitopes between H3 and H1 are 2.18, 0.63, 1.19, 2.43 and 1.90 Å, respectively.

The HA alignment of A/California/04/2009 with A/PR/8/1934 and of A/California/04/2009 with A/Aichi/2/1968 shows that neither A/PR/8/1934 nor A/Aichi/2/1968 has a gap-free alignment with A/California/04/2009. Thus, a new numbering scheme is required for the 2009 H1N1 ('swine flu') hemagglutinin. The new numbering starts at the same amino acid position as the H1 numbering [142, 143]. Amino acid 130 in A/California/04/2009 corresponds to a gap in A/PR/8/1934, indicating that amino acid position > 130 in A/California/ 04/2009 has the number equal to 1 plus the corresponding number of the amino acid in A/PR/8/1934. In Table 4.1, we use the A/California/04/2009 numbering scheme. HA alignment of swine flu strains deposited in NCBI and GISAID up until 18 May 2009 shows that no mutation occurred in amino acid 130. By examining the three-dimensional structure of A/PR/8/1934 (PDB code: 1RU7), we find that amino acid 129 (A/PR/8/1934 numbering) is exposed and position 130 is partially buried inside the molecule. Consequently, although amino acid 130 (A/California/04/2009 numbering) in swine flu HA may or may not be significant to antibody binding, we have no evidence that it is in the epitope.

4.2.2 Extension of mapped epitope using entropy method

The definition of information entropy of site k is

$$S(k) = -\sum_{i=1}^{20} \frac{n_i}{N} \ln \frac{n_i}{N}$$

where n_i is the number of times that amino acid i (i = 1–20) is found in site k of the aligned full-length strains. N is the number of those full-length sequences, equal to 2294.

The threshold of the information entropy values to identify epitopes is determined adaptively. Amino acids in epitopes are under immune pressure selection and evolve

to avoid recognition by antibodies, yielding large information entropy values for amino acids in epitopes among all H1 strains. Additionally, classical epitopes are defined to be on the surface of the three-dimensional structure of the hemagglutinin. We decreased the entropy threshold until amino acids inside the structure constitute a significant proportion of those amino acids newly incorporated into the epitopes. Here, we select 0.075 as the threshold.

4.2.3 Phylogenetic Tree and Its Root

Three hundred and twenty H1N1 swine flu protein sequences as of 18 May 2009 were downloaded from the NCBI and GISAID databases. ClustalW is applied to align these strains. Two hundred and sixty-six sequences containing residues 27324 (A/California/04/2009 numbering, modified from [3]) were used to create the phylogenic tree with PHYLIP [144]. Two hundred and sixty-six subsequences that included residues 27–324 were extracted, and duplicated subsequences were removed. The 23 unique aligned subsequences were then used as the input of the program PHYLIP. The distance matrix from protein sequences (protdist) and the Fitch-Margoliash and least squares methods with evolutionary clock (kitsch) [144] were sequentially applied to generate the phylogenic tree in Figure 4.1. The output of protdist is the input of kitsch. In the phylogenetic tree, the strain A/California/07/2009 (FJ966974) is one of the outgroups, which are closest to the root.

4.2.4 Calculation of p_{epitope}

The antigenic distance between two strains is calculated from the amino acid sequence in five epitopes of hemagglutinin. For each epitope, the P-value is defined as the proportion of different amino acids between these two strains. The largest of the

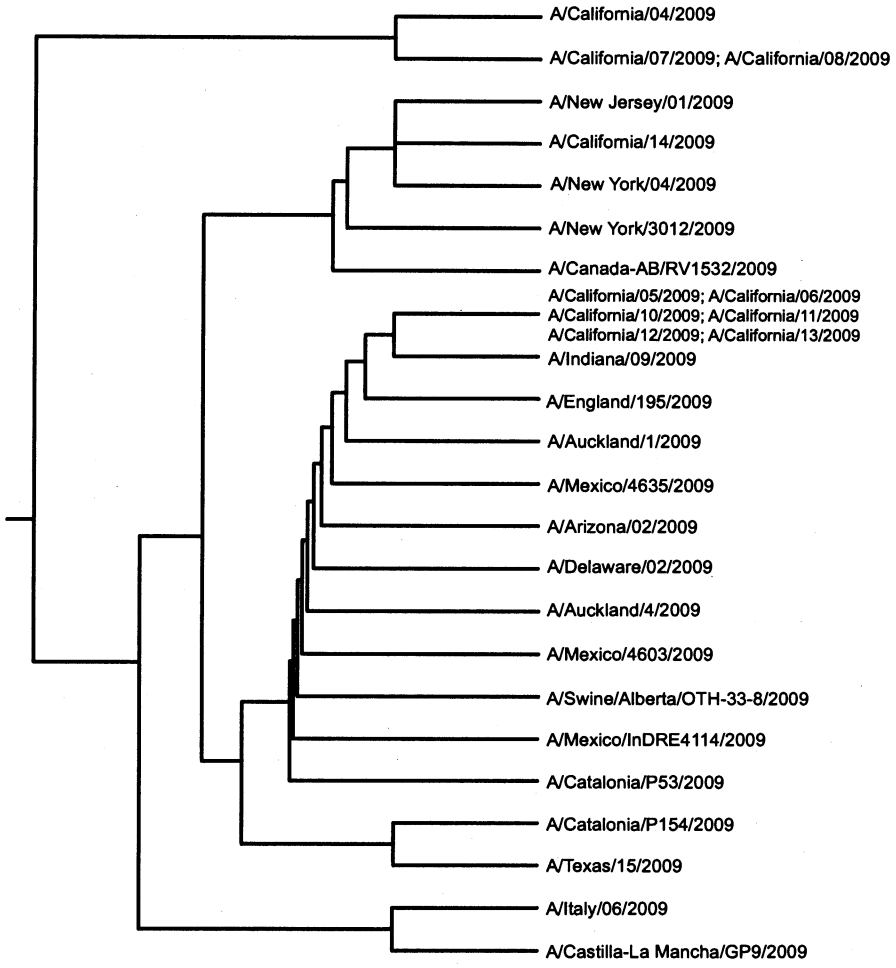


Figure 4.1 : Phylogenetic tree of swine flu hemagglutinins deposited in NCBI and GISAID until 18 May 2009. For each tip containing over two strains, representative strains are marked.

five P-values is defined as p_{epitope}, and the corresponding epitope is defined as the dominant epitope [2, 31].

4.3 Results

4.3.1 Antigenic Distance

As of 18 May 2009, there are 320 H1N1 swine flu strains in the NCBI and GISAID databases, and we focus on the interval covering residues 27–324 (A/California/04/2009 numbering, modified from [3]) in 266 of 320 sequences covering this interval. Among these 266 sequences, the closest strain to the root of a phylogenetic reconstruction [144] is A/California/07/2009 (FJ966974) (Figure 4.1). There are 20 mutations in the population of sequences, referenced to this strain, with a maximum Hamming distance of 4. Using the p_{epitope} method, we find the largest p_{epitope} value is 0.059 (dominant epitope = E). A/California/04/2009 (FJ966082) is a candidate for vaccine design. There are 20 mutations in the population of sequences, referenced to A/California/04/2009, with a maximum Hamming distance of 5. The largest p_{epitope} value between this candidate strain and all sequences deposited to date is 0.059 (dominant epitope = E), suggesting a worst-case vaccine efficacy against strains deposited to date of 69.1% of that of a perfect-match, $p_{\text{epitope}} = 0$, vaccine.

4.3.2 Epitope Identification

To construct p_{epitope} for H1N1, the identities of the epitope regions in H1 are needed. An early antibody mapping experiment on the PR/8 strain of H1N1 identified 9, 6, 6, 5 and 6 amino acids belonging to the five epitopes Sa, Sb, Ca1, Ca2 and Cb, respectively [3]. These H1 epitopes map by sequence homology to the H3 epitopes

B+D, B, C+D, A+D and E, respectively. Interestingly, for those H1 epitopes that map to multiple H3 epitopes, individual antibodies bound to all H1 epitopes mapping to the same set of H3 epitopes, suggesting that the identification of the H1 epitopes may be subject to refinement. The number of amino acids determined by antibody mapping to be in the five epitope residues of H1 is about one-third of the number of amino acids identified in the five epitopes of H3 influenza A [29, 30].

We determined the likely remaining members of the five epitope regions in H1 by information entropy methods. On 5 May 2009, we downloaded all 2735 H1 human influenza A strains from NCBI. We retained the 2294 full-length sequences. We constructed a sequence entropy diagram (Figure 4.2) from these data. The refined H1 epitopes were constructed by (i) mapping the known H3 epitopes [29, 30] to the H1 sequence, in rough agreement with the skeleton of sites identified by early antibody mapping experiments [3] (sequence mapping determined by ClustalW alignment of A/PR/8/1934 to A/Aichi/2/1968), and (ii) 31 additional sites identified as being under selective immune pressure, on the surface of the hemagglutinin protein, and with information entropy values > 0.075.

Interestingly, eight amino acids were identified with information entropy > 0.075 that were outside the conventional definition of epitopes. Three of these were in the tail region of the hemagglutinin (positions 3, 314 and 320, numbering as in [3]). The other five were not surface residues (positions 61, 174, 233, 248 and 249) (Figure 4.3). Although these residues can affect the geometry at the surface, and so can be under selective pressure, they are not available for presentation to antibodies and so cannot be within a classically defined antibody epitope.

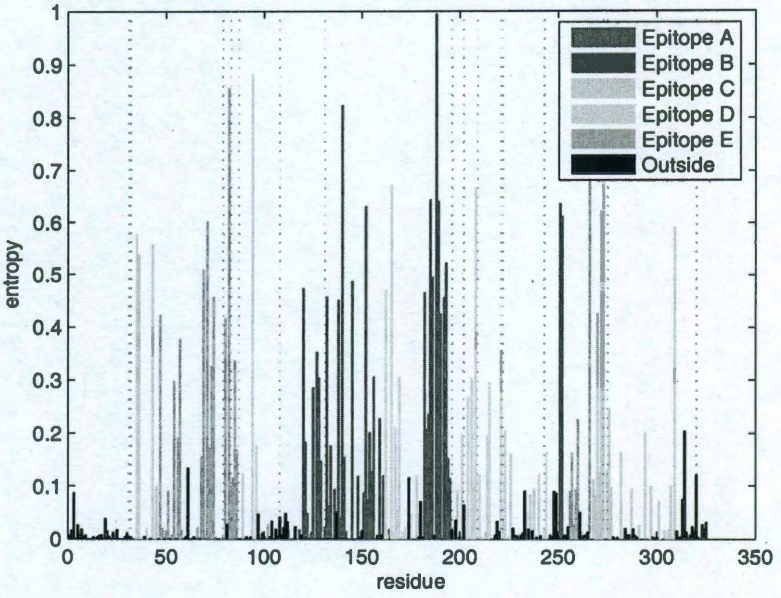


Figure 4.2 : Sequence entropy for the human strains of H1 (A/PR/8/1934 numbering, as in [3]). Positions belonging to predicted epitopes are color coded by the epitope identity.

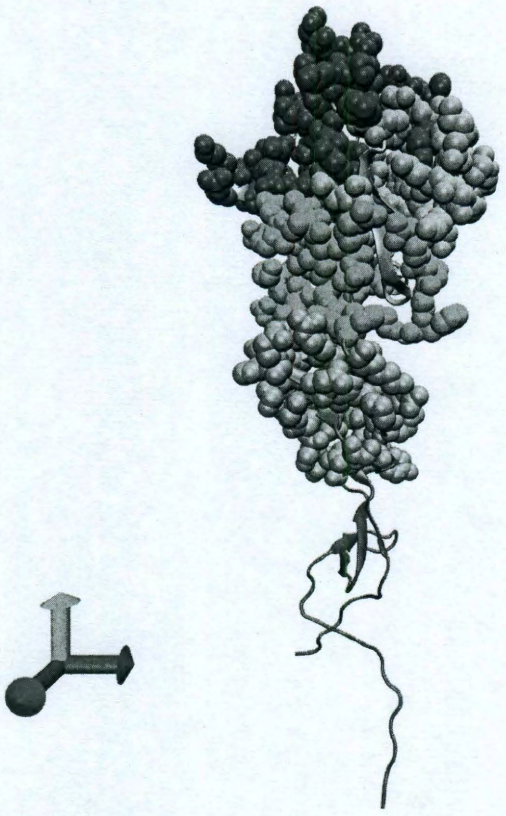


Figure 4.3 : Color-coded epitopes in the H1 structure (PDB code: 1RU7).

Table 4.1 : Amino acids in epitopes A, B, C, D and E of H1 (A/California/04/2009 numbering, modified from [3]). For A/PR/8/1934 numbering, amino acid numbers above 130 would have 1 subtracted from them.

Epitope	Amino acids
A	118 120 121 122 126 127 128 129 132 133 134
	135 137 139 140 141 142 143 146 147 149 165
	252 253
B	124 125 152 153 154 155 156 157 160 162 183
	184 185 186 187 189 190 191 193 194 195 196
C	34 35 36 37 38 40 41 43 44 45 269 270 271 272
	273 274 276 277 278 283 288 292 295 297 298
	302 303 305 306 307 308 309 310
D	89 94 95 96 113 117 163 164 166 167 168 169
	170 171 172 173 174 176 179 198 200 202 204
	205 206 207 208 209 210 211 212 213 214 215
	216 222 223 224 225 226 227 235 237 239 241
	243 244 245
E	47 48 50 51 53 54 56 57 58 66 68 69 70 71 72 73
	74 75 78 79 80 82 83 84 85 86 102 257 258 259
	260 261 263 267

4.4 Discussion

With the epitopes in H1 identified, p_{epitope} can be constructed for H1 influenza A. The parameter p_{epitope} provides a quantitative definition of antigenic distance. With this measure of antigenic distance, vaccine strains as 'close' as possible to potential circulating strains can be identified. This capability should be useful in the design of the H1N1 component of the annual influenza vaccine. This capability should also be useful for special situation H1N1 vaccines, such as a vaccine for the recently emerged swine flu.

The p_{epitope} measure of antigenic distance can be used to estimate vaccine efficacy. Vaccine efficacy is $(u - v)/u$, where u is the probability (or rate) that unvaccinated people are infected and v the probability (or rate) that vaccinated people are infected. By analogy with the H3N2 study [2, 31], we expect vaccine efficacy will be well predicted by the equation $E = 0.47 - 2.47 \times pepitope$. This equation predicts, for example, that a single mutation in dominant epitope B would lead to a vaccine efficacy that is 76.1% of that of a perfect match between vaccine and virus. This equation also predicts that vaccine efficacy is no longer positive for $p_{\text{epitope}} > 0.19$.

Chapter 5

Quantifying Selection and Diversity in Viruses by Entropy Methods, with Application to the Hemagglutinin of H3N2 Influenza

Many viruses evolve rapidly. For example, hemagglutinin of the H3N2 influenza A virus evolves to escape antibody binding. This evolution of the H3N2 virus means that people who have previously been exposed to an influenza strain may be infected by a newly emerged virus. In this chapter, we use Shannon entropy and relative entropy to measure the diversity and selection pressure by antibody in each amino acid site of H3 hemagglutinin between the 1992–1993 season and the 2009–2010 season. Shannon entropy and relative entropy are two independent state variables that we use to characterize H3N2 evolution. The entropy method estimates future H3N2 evolution and migration using currently available H3 hemagglutinin sequences. First, we show that the rate of evolution increases with the virus diversity in the current season. The Shannon entropy of the sequence in the current season predicts relative entropy between sequences in the current season and those in the next season. Second, a global migration pattern of H3N2 is assembled by comparing the relative entropy flows of sequences sampled in China, Japan, the USA, and Europe. We verify this entropy method by describing two aspects of historical H3N2 evolution. First, we identify 54 amino acid sites in hemagglutinin that have evolved in the past to evade the immune system. Second, the entropy method shows that epitopes A and B in the top of hemagglutinin evolve most vigorously to escape antibody binding. Our

work provides a novel entropy-based method to predict and quantify future H3N2 evolution and to describe the evolutionary history of H3N2.

5.1 Introduction

A common strategy by which viruses evade pressure from the immune system is to evolve and change their antigenic profile. Viruses with a low evolutionary rate that infect only humans, such as the small pox virus [35], can be effectively controlled by vaccinating the human population. By contrast, viruses with a high evolutionary rate, such as HIV, hepatitis B, and influenza A, resist being eliminated by the immune system by generating a plethora of mutated virus particles and causing chronic or repeated infection. In this study, we take subtype H3N2 influenza A virus as a model evolving virus. Influenza A virus circulates in the human population every year, typically causing 3–5 million severe illnesses and 250,000–500,000 fatalities all over the world [14]. Hemagglutinin (HA) and neuraminidase (NA) are two kinds of virus surface glycoproteins encoded by the influenza genome. The subtype of influenza is jointly determined by the type of hemagglutinin ranging from H1 to H16, and that of neuraminidase ranging from N1 to N9. On the surface of the virus membrane, HA exists as a cylindrical trimer containing three HA monomers, and each monomer comprises two domains, HA1 and HA2. Hemagglutinin is also a key factor in virus evolution, because it is the major target of antibodies, and HA escape mutation changes the antigenic character of the virus presented to the immune system. The H3N2 virus causes the largest fraction of influenza illness. H3 hemagglutinin is under selection by the immune response mainly on the five epitope regions in the HA1 domain [36], labeled epitopes A to E, as shown in Figure 5.1. The immune pressure and the escape mutation drive the evolution of the H3N2 virus. The underlying

mutation rate of the HA gene is 1.6×10^{-5}/amino acid position/day [37], measured using the method modified from that in an earlier study on the HA mutation rate [38]. Note that the mutation rate does not necessarily equal to the evolution rate, or the fixation rate. The mutation rate equals to the evolutionary rate only if the evolution is neutral. The non-neutrality of the HA evolution is shown in the Results section. Evolution of the hemagglutinin viral protein causes occasional mismatch between the virus and the vaccine and decreases vaccine effectiveness [2, 34]. As more amino acid substitutions are introduced into influenza sequences, the antigenic characteristics of influenza strains drift away [39], and influenza epidemic severity of subtype H1N1 [40] and subtype H3N2 [41] increases.

The H3N2 virus has a distinguished evolutionary history, largely affected by the immune pressure. The H3N2 virus emerged in the human population in 1968 and has been circulating in the population since 1968. The phylogenetic tree of H3 hemagglutinin since 1968 has a linear topology in which most sequences are close to the single trunk of the tree, and the lengths of the branches are short [42, 30, 43]. Historical hemagglutinin sequences fall into a series of clusters, each of which has similar genetic or antigenic features and circulates for 2–8 years before being replaced by the next cluster [44, 30]. The evolution of different amino acid positions of hemagglutinin shows a remarkable heterogeneity: a subset of positions undergo frequent change, while some positions are conserved [4]. This heterogeneity is quantified by the Shannon entropy at each position of the amino acid sequence of hemagglutinin [1]. Shannon entropy has been used to locate protein regions with high diversity, such as the antigen binding sites of T-cell receptors [45]. Shannon entropy has been used to identify antibody binding sites, or epitopes, which are under immune pressure and so are rapidly evolving [1]. The heterogeneity of amino acid substitution suggests that

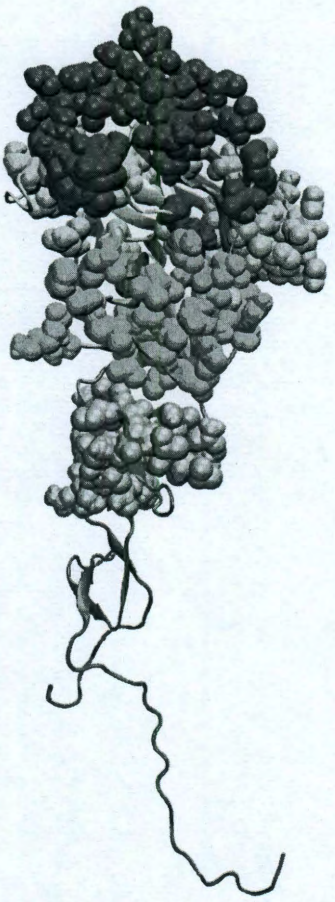


Figure 5.1 : The tertiary structure of the HA1 domain of H3 hemagglutinin (PDB code: 1HGF). The surface of HA1 facing outward is the exposed surface when the hemagglutinin trimer is formed. The other two HA1 domains (not shown) in the HA trimer are located at the back of the structure displayed here. The solid balls represent five epitopes. Color code: blue is epitope A, red is epitope B, cyan is epitope C, yellow is epitope D, and green is epitope E.

point mutations randomly occurring in distinct positions have different contributions to the virus fitness.

The selection pressure on the H3N2 virus to evolve is reflected in the difference between the H3 hemagglutinin sequences in two consecutive seasons. We consider Northern hemisphere strains. When the epidemic initiates in a new season, we assume that each position of an HA sequence inherits the amino acid from a sequence of the previous season or has a different amino acid due to random mutation and selection. This assumption comes from the fact that the H3N2 virus circulating in each influenza season migrates from a certain geographic region in which the virus is preserved between two influenza seasons [43, 32]. In the absence of selection, the histogram of the 20 amino acid usage in one position in the current season is similar to that in the same position in the previous season except for changes due to the small random mutation rate. The difference between the two histograms beyond that expected due to mutation quantifies selection.

Synthesizing these factors, we introduce an entropy method to describe the evolution of influenza. The entropy method extracts an evolutionary pattern from aligned sequences. Shannon entropy quantifies the amount of sequence information in each position of aligned sequences [46, 47]. The sequence information reflects the variation, which is equivalently diversity, in each position, and so Shannon entropy has been used to measure the diversity in each position [48, 49, 50]. Shannon entropy has also been used to measure the structural conservation in the protein folding dynamics [51, 52]. See [53] for a detailed review of the applications of Shannon entropy. On the other hand, relative entropy measures gain of sequence information at each position and requires a background amino acid frequency distribution [54]. Relative entropy was also used as a sequence conservation measure to detect functional pro-

tein sites [55, 56]. Further, a dimension reduction technique using relative entropy has identified sectors in proteins [57, 56]. As an extension of these previous works, we apply Shannon entropy and relative entropy to jointly measure two quantities in each position: sequence information in one season and gain of sequence information from one season to the next season. Simultaneous analysis of Shannon entropy and relative entropy sheds light on the evolutionary pattern of the H3N2 virus evolution when data from multiple seasons are available. In the HA1 domain, positions in the epitope regions have increased Shannon entropy, and this feature was applied to locate the epitopes of H1 hemagglutinin [1]. We here use Shannon entropy to quantify the virus diversity in each amino acid position in each season. The entropy relative to the previous season [58] is also used to analyze the evolution of the HA1 domain in one single season and to quantify the selection pressure on the virus in each amino acid position in each season. The selection and the virus diversity are two significant state variables determining the dynamics of evolution.

The chapter is organized as follows. The Materials and Methods section presents the data used in this work and details of the entropy model. The Results section uses Shannon entropy of the sequence in one season to predict the evolution quantified by relative entropy from this season to the next season. Results are also presented for the flow of virus migration between the four geographic regions of China, Japan, the USA, and Europe. In the Result section, we demonstrate the entropy method in two applications, the results of which agree with prior knowledge on H3N2 evolution. Finally, we discuss our results and present our conclusions.

5.2 Materials and Methods

5.2.1 Sequence Data

The hemagglutinin sequences considered in this work are the amino acid sequences of the hemagglutinin of human influenza A H3N2 virus. We only use Northern Hemisphere sequences because 90% of the world population lives in Northern Hemisphere. The influenza season in Northern Hemisphere is defined as the time interval from October in one year to September in the next year. We downloaded 5309 Northern Hemisphere sequences labeled with month of collection from the NCBI Influenza Virus Resource on 16 January 2011. Sequences too short to cover residues 1–329 of hemagglutinin were removed, and the remaining sequences were trimmed to only keep residues 1–329 in the HA1 domain. Any sequence with an undefined amino acid in residues 1–329 was removed. We consider 18 seasons from 1992–1993 to 2009–2010 during which most available sequences were collected. In total, we preserved and aligned 4292 sequences in these 18 seasons containing amino acids 1–329.

5.2.2 Histograms of 20 Amino Acids

The first step is to quantify the alignment of the amino acid sequences. The aligned historical H3 hemagglutinin sequences form a matrix $\mathbf{A}$ with 4292 rows and 329 columns. The element $\mathbf{A}_{l,j}$ denotes the identity of the amino acid in sequence l and position j. The 4292 sequences were clustered into 18 groups by the seasons of sampling from the 1992–1993 season to the 2009–2010 season. Note that most of the sequences before the 1992–1993 season were not labeled with month of collection and are excluded from this study. We denote by $i = 0, 1, \ldots, 17$ the seasons between 1992–1993 and 2009–2010. For position j in season i, the relative frequency of each

amino acid k, $f(k,i,j), k = 1,\ldots,20$, was calculated from the vector $\mathbf{A}_{\vec{l}(i),j}$ where the index array $\vec{l}(i)$ holds the indices of sequences sampled in season i.

5.2.3 Shannon Entropy as Diversity

The Shannon entropy is one useful quantification of the diversity in single position. Large Shannon entropy has the physical meaning that the amino acid in the given position is prone to be substituted. This physical meaning was also applied in [1]. The diversity at a single position takes the format of Shannon entropy because of the sensitivity of Shannon entropy to diversity.

This physical meaning of the Shannon entropy does not necessarily involve the joint frequency distributions for two and more positions, and we do not consider the joint frequency in the manuscript. Rather, we define the diversity only in the level of single amino acid position. Consequently, the defined diversity is additive for a number of positions. The idea of adding diversity in each position of the sequence comes from classic works such as [46], which added the Shannon entropy in each position to measure the total diversity in an aligned binding site.

Therefore, diversity of the virus in each position in each season is represented by the Shannon entropy that quantifies the amount of information in the histogram or distribution under study. For the sequences sampled in all the seasons, positions with high evolutionary rate have a higher Shannon entropy compared to the conserved positions [1]. The sequences in each season are assumed to be collected concurrently. The Shannon entropy is a quantification of diversity of amino acids in one position, and so the diversity in position j in season i is calculated from the histogram $\mathbf{f}(i,j) =$

$[f(1,i,j),\ldots,f(20,i,j)]^{\mathrm{T}}$ by Shannon entropy

$$D_{i,j} = -\sum_{k=1}^{20} f(k,i,j)\log f(k,i,j) \tag{5.1}$$

in which $k = 1,\ldots,20$ is the identity of the amino acid in position j in season i.

5.2.4 Relative Entropy as Selection Pressure

Selection in each position j in season i is reflected by the significant difference between the 20-bin histogram in the current season $\mathbf{f}(i,j)$ and that in the previous season $\mathbf{f}(i-1,j)$. In the absence of selection, random mutation and genetic drift are the dominant forces generating $\mathbf{f}(i,j)$ from $\mathbf{f}(i-1,j)$. In each position, random mutation creates a slightly modified histogram, from which amino acids are randomly chosen to appear in season i by the effect of genetic drift.

The source of random mutation is the spontaneous error of the RNA polymerase replicating the influenza virus RNA. The random mutation rate in different regions of hemagglutinin is thought to be homogeneous, regardless whether the regions are in antigenic sites or not [145]. Therefore random mutation is modeled as a Poisson process $\mathbf{M}(\mu, g(k))$ equally affecting all the positions. Here μ is the mutation rate of influenza A virus that equals to 5.8×10^{-3}/residue/season [37], and $g(k), k = 1,\ldots,20$ is the relative frequency of each amino acid in the whole alignment A. The probability that the original amino acid t mutates to amino acid u is

$$\mathbf{M}_{u,t}(\mu, g) = \frac{\mu g(u)}{1-g(t)}. \tag{5.2}$$

So after mutating for one season, the histogram in position j in season $i-1$ is obtained by

$$\hat{\mathbf{f}}(i,j) = \mathbf{M}(\mu, g)\,\mathbf{f}(i-1,j). \tag{5.3}$$

This histogram serves as the background distribution for season i from which the sequences in season i are built.

If selection is absent, the effect of genetic drift is to create sequences in the current season by randomly choosing amino acids in each position from a background distribution $\hat{\mathbf{f}}(i,j)$. We denote by N_i the number of sequence in season i. The probability that N_i amino acids in position j have the histogram $\mathbf{f}(i,j)$ is [56]

$$\Pr\{\mathbf{f}(i,j)\} \approx \exp(-N_i S_{i,j}) \tag{5.4}$$

where

$$S_{i,j} = \sum_{k=1}^{20} f(k,i,j) \log \frac{f(k,i,j)}{\hat{f}(k,i,j)} \tag{5.5}$$

is the relative entropy between the observed histogram, $f(k,i,j)$, and the background histogram, $\hat{f}(k,i,j)$ [58].

The null hypothesis that selection is absent in the evolution is rejected if the relative entropy $S_{i,j}$ is great enough such that the probability in Equation 5.4 is less than 0.05, that is, the relative entropy $S_{i,j}$ is greater than $-\log(0.05)/N_i$ in season i. Note that the majority of residues were stable in most of the seasons, and in this case the relative entropy is $S_{i,j} = \log(1/(1-\mu)) \approx \mu$. To avoid classifying these stable residues erroneously as positions under selection, a proper threshold of relative entropy needs to be larger than the mutation rate μ. Additionally, a fraction λ of the circulating HA1 sequences were not deposited in the database because of the sampling bias of the HA1 sequences. In an extreme case, in a stable position j with the real histogram of 20 amino acids $[1-\lambda, \lambda, \ldots, 0]^{\mathrm{T}}$ in all the seasons, the histograms of the sequences sampled in two consecutive seasons $i-1$ and i are $\mathbf{f}(i-1,j) = [1,0,\ldots,0]^{\mathrm{T}}$ and $\mathbf{f}(i,j) = [1-\lambda, \lambda, \ldots, 0]^{\mathrm{T}}$, respectively, and so the relative entropy introduced

by the sampling bias is

$$S^{\text{bias}} \approx (1-\lambda)\log\frac{1-\lambda}{1-\mu} + \lambda\log\frac{\lambda}{\mu/19} \tag{5.6}$$

in spite of the absence of selection in position j in season i. The relative entropy S^{bias} equals to 0.1 if a sampling bias $\lambda = 2.5\%$ exists in the HA1 database sequences. We fix the threshold of the relative entropy in season i to

$$S_i^{\text{thres}} = \max\left\{-\log(0.05)/N_i, (1-\lambda)\log\frac{1-\lambda}{1-\mu} + \lambda\log\frac{\lambda}{\mu/19}\right\} \approx \max\{3/N_i, 0.1\}. \tag{5.7}$$

The numbers of collected HA1 sequences N_i were fewer than 30 only in the 1995–1996 season ($i = 3$) with $N_3 = 25$. The thresholds $S_3^{\text{thres}} = 0.12 > 0.1$ in the 1995–1996 season due to the small numbers of HA1 sequences. In all the other 17 seasons, the numbers of sequences N_i were greater than 30, and so the thresholds $S_i^{\text{thres}} = 0.1$.

5.3 Results

In this section, we show the positive correlation between the Shannon entropy in season i and the relative entropy from season i to season $i+1$. This correlation means that the larger the virus diversity in one season, the higher the virus evolutionary rate from this season to the next season. We draw the H3N2 migration pattern by comparing the relative entropy. The migration pattern reveals a novel migration path from the USA to Europe and shows that the virus evolutionary rate is higher in the epicenter, China, than in the migration paths. We also demonstrate the entropy method in two applications. First, we compute average Shannon entropy and relative entropy in each position over the past 17 seasons to identify positions under selection pressure. Second, we compare Shannon entropy and relative entropy in epitope regions to find the contribution of each epitope to the H3N2 evolution. Results of these

two applications agree with previous studies and additionally show the heterogeneity of the selection pressure over different amino acid positions of hemagglutinin, with increased pressure in the epitopes, as well as the dominance of epitopes A and B.

5.3.1 Correlation between Shannon entropy and relative entropy

Relative entropy $S_{i+1,j}$ in amino acid position j from season i to season $i+1$ linearly increases with Shannon entropy $D_{i,j}$ in position j in season i. For the sequences sampled from the 1992–1993 season ($i = 0$) to the 2008–2009 season ($i = 16$), $329 \times 17 = 5593$ ordered pairs $(D_{i,j}, S_{i+1,j})$ are calculated. All except two pairs fall into 8 bins in which the values of $D_{i,j}$ belong to 8 intervals $[0, 0.1)$, $[0.1, 0.2)$, $\ldots$, $[0.7, 0.8)$, respectively. The first bin with $D_{i,j}$ in $[0, 0.1)$ is discarded because it contains numerous conserved amino acid positions. The values of $S_{i+1,j}$ are averaged respectively in each of the 7 remaining bins. As described in Figure 5.2, average relative entropy in each bin shows positive correlation with midpoints of the $D_{i,j}$ interval, $R^2 = 0.70$. An amino acid position j with high Shannon entropy $D_{i,j}$ in season i is expected to present high relative entropy $S_{i+1,j}$ from season i to $i+1$. The evolution in position j from season i to $i+1$ quantified by relative entropy is therefore predicted using the mean and standard error of $S_{i+1,j}$ in the bin chosen by $D_{i,j}$ in season i.

Positive correlation is also observed between the mean values of $D_{i,j}$ and $S_{i+1,j}$ in a variety of positions j in each season i. In each season between 1992–1993 ($i = 0$) and 2008–2009 ($i = 16$), we average the Shannon entropy $D_{i,j}$ and relative entropy $S_{i+1,j}$ over the positions j with $D_{i,j} > 0.1$. The data point with $i = 1$, $(0.22, 1.38)$, has a large standard error of the relative entropy $S_{i+1,j}$ and is excluded in the analysis below. The remaining average Shannon entropy $\langle D \rangle_i$ ($i = 0, 2, \ldots, 16$) correlates with average relative entropy $\langle S \rangle_{i+1}$ ($i = 0, 2, \ldots, 16$) with $R^2 = 0.50$, as shown in Figure

5.3. A least squares fit gives $\langle S\rangle_{i+1} = 1.82\langle D\rangle_i - 0.23$. Thus, the expected average relative entropy from the current season i to the next season $i+1$ can be calculated from the average Shannon entropy $\langle D\rangle_i$ in the current season i.

The above relationships between Shannon entropy and relative entropy cannot be generated by a neutral evolution model. To demonstrate this result, we create an ensemble of 1000 identical sequences with 50 amino acid positions. Each iteration of the model simulates H3N2 evolution during one season. In each iteration, the number of mutated amino acids N_{mut} in each sequence follows a Poisson distribution with mean $\mu = 2.0$, which is the annual substitution rate in history. The N_{mut} mutated positions are then randomly assigned in the corresponding sequence. We randomly select $p_{\text{cut}} = 10\%$ of the sequences to build the sequence ensemble in the next iteration. The Shannon entropy $D_{i,j}$ and relative entropy $S_{i+1,j}$ generated in iteration $i = 51$–100 are processed using the same method as for H3 sequences in history. First, as shown in Figure 5.2, no increasing trend appears in the means of $S_{i+1,j}$ in the 7 bins from $[0.1, 0.2)$ to $[0.7, 0.8)$. Second, no correlation ($R^2 = 0.003$) is observed between $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$, see Figure 5.3. When we change the parameters μ between 1.0 and 10 and p_{cut} between 1% and 100% in the algorithm, the simulation still does not yield the visible increasing trend in Figure 5.2 or the correlation observed in Figure 5.3. As a result, we conclude that neutral evolution alone is not able to generate the pattern between Shannon entropy and relation entropy of H3 sequences. It was previously shown that the fixation rate of H3N2 evolution cannot be explained only by neutral evolution [4]. In this study, the monotonically increasing linear relationship between relative entropy and Shannon entropy in Figure 5.2 and 5.3 suggests that selection pressure substantially contributes to H3N2 evolution.

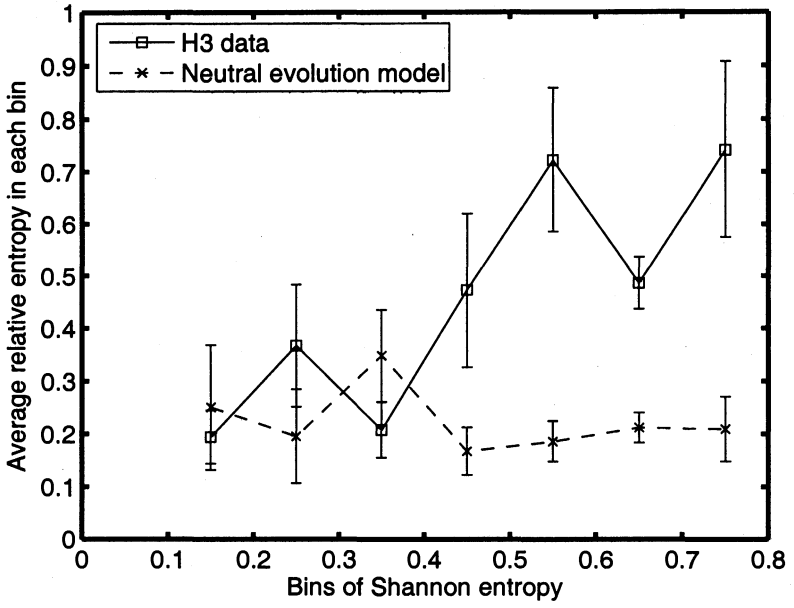


Figure 5.2 : Mean and standard error of relative entropy $S_{i+1,j}$ in each bin of Shannon entropy. Shannon entropy and relative entropy in each of the 329 positions and in each of the 17 seasons between 1992–1993 ($i = 0$) and 2008–2009 ($i = 16$) fall into one of the eight bins. The first bin with Shannon entropy less than 0.1 is discarded. Bins with larger Shannon entropy $D_{i,j}$ also have larger relative entropy $S_{i+1,j}$. Shannon entropy $D_{i,j}$ and relative entropy $S_{i+1,j}$ in iteration $i = 51$–100 of the neutral evolution model are used to calculate mean and standard error of relative entropy in each bin of Shannon entropy distribution in the same way. No increasing trend is found. Error bar is one standard error.

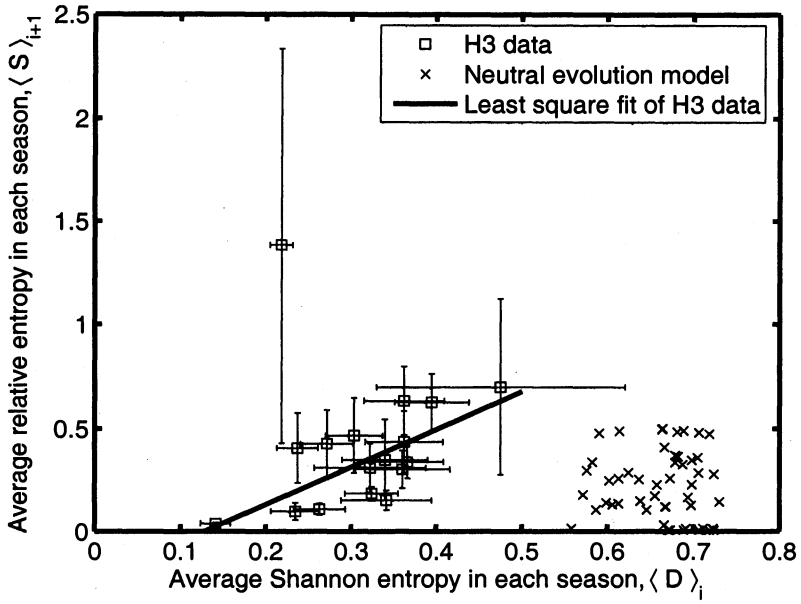


Figure 5.3 : Average Shannon entropy $\langle D\rangle_i$ versus average relative entropy $\langle S\rangle_{i+1}$ for each season between 1992–1993 ($i = 0$) and 2008–2009 ($i = 16$). For each season i, a set of amino acid positions j with Shannon entropy $D_{i,j}$ greater than 0.1 are chosen. For all the j in this set of positions, $\langle D\rangle_i$ is the average of the Shannon entropy $D_{i,j}$ values and $\langle S\rangle_{i+1}$ is the average of relative entropy $S_{i+1,j}$ values. Horizontal and vertical error bars are the standard errors of Shannon entropy and relative entropy, respectively. The solid line, $\langle S\rangle_{i+1} = 1.82\langle D\rangle_i - 0.23$, is a least squares fit of $\langle D\rangle_i$ to $\langle S\rangle_{i+1}$ ($i = 0, 2, \ldots, 16$). A strong correlation with $R^2 = 0.50$ exists between $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ excluding the point $(0.22, 1.38)$ with $N_i = 1$, which has a large standard error of the relative entropy $S_{i+1,j}$. Using the same method, $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ are calculated from a neutral evolution model, $i = 51$–100, and plotted. No visible correlation exists between $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ from the neutral evolution model.

5.3.2 Annual Virus Migration

The entropy method is also used to analyze the global migration pattern of the virus. Most of the Northern Hemisphere H3 sequences were collected in East-Southeast Asia, the USA, and Europe. East-Southeast Asia is suggested to be the reservoir of the annual H3N2 epidemic [43]. To increase the geographic resolution in East-Southeast Asia, we use two regions, China and Japan, as the representative of East-Southeast Asia, because each of these regions has a population over 50 million and has a consistent time series of H3 sequence data from the 2001–2002 to the 2007–2008 season.

We select the sequences in four geographic regions that are China, Japan, the USA, and Europe in seven seasons from 2001–2002 to 2007–2008. In all the six pairs of consecutive seasons, we calculate for each region four average relative entropy values of the whole sequence. These four average relative entropy values are calculated using the sequences in each of the four regions in the previous season as the reference. The results are shown in Table 5.1. Sequences collected in China in the previous season yield the minimum relative entropy to the sequences in the current season collected in China ($p < 2.1 \times 10^{-5}$, Wilcoxon signed-rank test), Japan ($p < 0.0049$, Wilcoxon signed-rank test), and the USA ($p < 0.0012$, Wilcoxon signed-rank test). Sequences in the USA in the previous season have the minimum relative entropy to the sequences in Europe in the current season ($p < 0.15$, Wilcoxon signed-rank test). Relative entropy data in Table 5.1 imply the virus migration from China, as the geographic reservoir, to Japan and the USA and suggest a migration from the USA to Europe. The result in Table 5.1 also implies that the H3N2 virus circulating in China seed the virus in China, Japan, and the USA in the next season, and the virus in the USA probably seed the virus in Europe in the next season.

Table 5.1 : The relative entropy between hemagglutinin sequences in the different regions in the current influenza season and sequences in these regions in the previous season. The minimum relative entropy in each column is marked in bold. The p values of the Wilcoxon signed-rank test between the minimum relative entropy and other relative entropy values in the same column are in the parentheses. Hemagglutinin sequences were collected from four geographic regions: China, Japan, the USA, and Europe. Seven seasons from 2001–2002 to 2007–2008 are used here. The relative entropy values listed in this table are averaged for all the sites and all the six pairs of consecutive seasons. These results imply that the H3N2 viruses in China, Japan, and the USA migrate from China, while the H3N2 virus in Europe migrates from USA.

Region of the current season / Region of the previous season	China	Japan	USA	Europe
China	**0.057**	**0.040**	**0.044**	0.064 (0.0017)
Japan	0.114 (2.1×10^{-5})	0.094 (0.0049)	0.076 (0.0012)	0.059 (0.032)
USA	0.105 (2.1×10^{-10})	0.087 (3.3×10^{-8})	0.070 (6.4×10^{-5})	**0.056**
Europe	0.135 (3.8×10^{-9})	0.115 (3.4×10^{-6})	0.094 (4.8×10^{-7})	0.074 (0.15)

Comparison of the relative entropy data in Table 5.1 in the H3N2 reservoir and migration paths also reveals the H3N2 virus migration pattern. Using the Wilcoxon sign-rank test, the relative entropy data of the H3 hemagglutinin in China in two consecutive seasons is significantly greater than those in the three migration paths: from China to Japan ($p = 0.035$), from China to the USA ($p = 0.0030$), and from the USA to Europe ($p = 0.0017$). The relative entropy data, therefore, confirms China as the H3N2 reservoir and implies that novel H3N2 viruses are emerging in China, not during the migration process.

5.3.3 Positions under Selection

The values of diversity as Shannon entropy $D_{i,j}$ and selection as relative entropy $S_{i,j}$ are available for the sequences collected from the 1993–1994 season to the 2009–2010 season. First we apply the mean field approximation to remove the variation of selection and diversity over the time, and only consider the variation of Shannon entropy and relative entropy in different positions and regions over the past 17 seasons. A profile of the pattern of Shannon entropy and relative entropy in position j comprises the average selection, the number of seasons under selection, and the average diversity. The average selection $\bar{S}_j$ is expressed by the mean of relative entropy in each position over the 17 seasons

$$\bar{S}_j = \frac{1}{17}\sum_{i=1}^{17} S_{i,j} \tag{5.8}$$

and is displayed in Figure 5.4 (a). The number of seasons under selection in each position j is calculated by

$$N_j = \sum_{i=1}^{17} H\left(S_{i,j} - S_i^{\text{thres}}\right) \tag{5.9}$$

where H is the Heaviside step function. The numbers are shown in Figure 5.4 (b). The average diversity $\bar{D}_j$ in each position is calculated by averaging the Shannon entropy over the 17 seasons from 1993–1994 to 2009–2010

$$\bar{D}_j = \frac{1}{17}\sum_{i=1}^{17} D_{i,j} \tag{5.10}$$

and is displayed in Figure 5.4 (c).

Figure 5.4 (d) presents the distribution of the selection. Around 76% of the amino acid positions 1–329 of hemagglutinin have an average selection close to zero and fall into the leftmost bin. The numbers of seasons when selection $S_{i,j}$ in these positions were greater than the threshold level S_i^{thres} are shown in Figure 5.4 (e). The average diversities $\bar{D}_{i,j}$ in all the positions are shown in Figure 5.4 (f). If position j is under selection with $S_{i,j} > S_i^{\text{thres}}$ in greater than two of the 17 seasons between 1993–1994 and 2009–2010, or $N_j > 2$, this position j is classified as a position under selection in the evolutionary history of H3N2 virus. The 54 positions with $S_{i,j} > S_i^{\text{thres}}$ in greater than two seasons are listed in Table 5.2.

Patterns of selection and diversity similar to those observed in historical sequences in Figure 5.4 are generated by a Monte Carlo simulation model, as displayed in Figure S1. The basis of the Monte Carlo simulation is that antibody binds to one of two epitope regions on the surface of the HA1 domain [2], and the dominant epitope bound by antibody is under immune pressure and undergoes a higher substitution rate [42]. The detailed description and discussion of the Monte Carlo model is in the Appendix.

5.3.4 Comparison of Different Regions

A human antibody binds to five epitopes in the H3 hemagglutinin [36]. The five epitopes are located in different parts of the HA1 domain of the cylinder-like structure

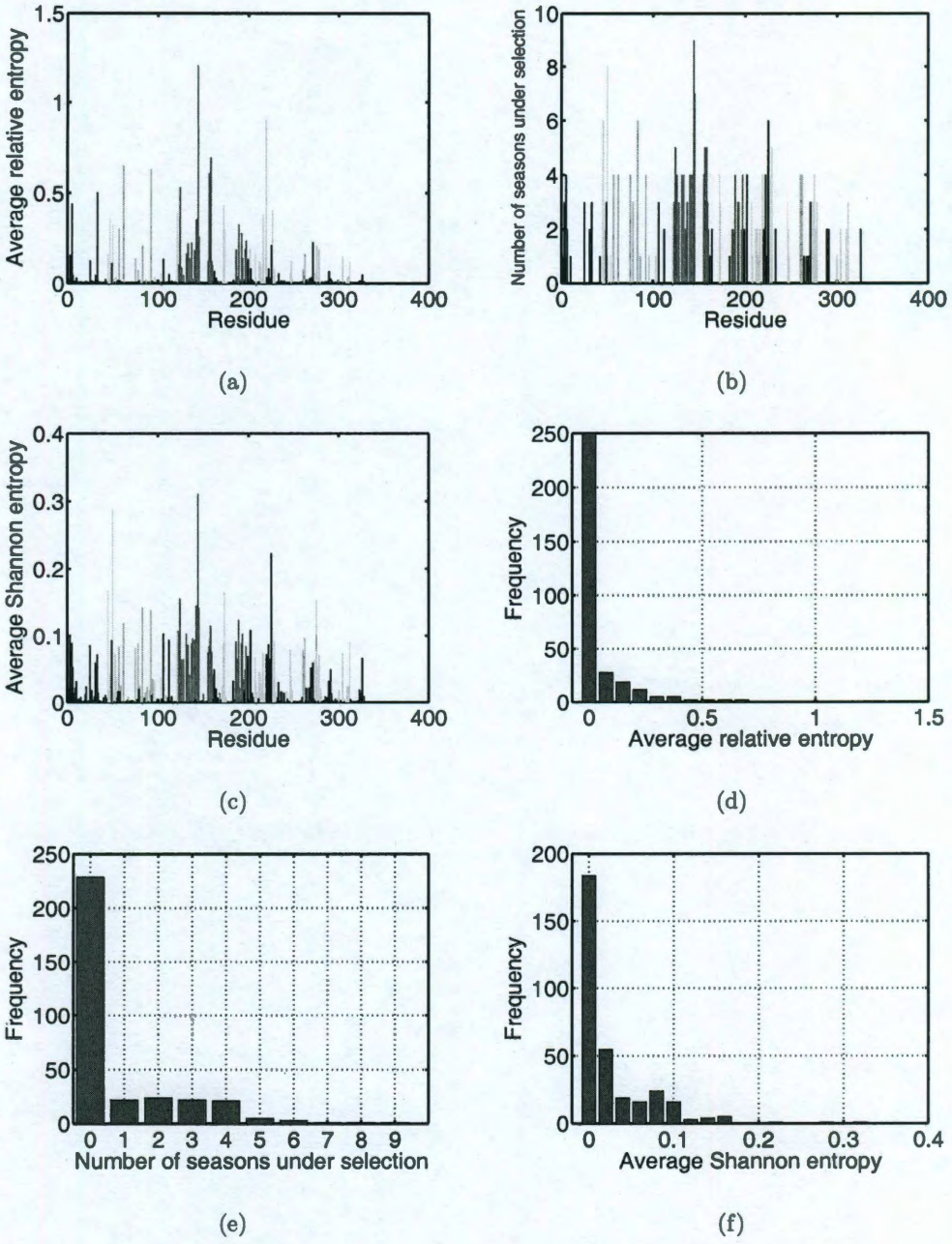


Figure 5.4 : (a) Average selection in each position quantified by relative entropy during the past 17 seasons from 1993–1994 to 2009–2010, calculated by $\bar{S}_j = \sum_{i=1}^{17} S_{i,j}/17$. The colors represent positions in epitopes A to E and positions outside the epitopes, as in Figure 5.1. (b) Number of seasons for each position when the relative entropy was greater than the threshold S_i^{thres}, i.e. the position was under selection. (c) Average diversity in each position quantified by Shannon entropy in the seasons from 1993–1994 to 2009–2010, calculated by $\bar{D}_j = \sum_{i=1}^{17} D_{i,j}/17$. (d) Distribution of the average selection in each position displayed in (a). (e) Distribution of the numbers of seasons under selection displayed in (b). (f) Distribution of the average diversity in each position shown in (c).

Table 5.2 : Amino acid positions j under selection. To be included, the positions must be under selection, $S_{i,j} > S_i^{\text{thres}}$, in greater than two seasons.

Region	Amino acid positions
Epitope A	122 124 126 131 133 137 140 142 144 145
Epitope B	128 155 156 157 158 159 189 192 193 197
Epitope C	45 50 273 275 278 312
Epitope D	121 172 173 201 207 219 226 227 229 246
Epitope E	57 62 75 78 83 92 260 262
Out of epitopes	3 5 25 33 49 106 202 222 225 271

of the H3 hemagglutinin. Epitopes A and B are on the top of the HA1 structure and are exposed in the HA trimer. Epitope D is on the top of HA1 structure and is partly buried inside the HA trimer. Epitopes C and E are at the central area of the exposed surface of the HA1 domain as shown in Figure 5.1. Using the entropy method, we will show that epitopes A and B are under the highest average selection over all the seasons. These results can be interpreted as the antibody binds mostly to the top exposed part of the structure of hemagglutinin trimer defined by epitopes A and B, and so the selection in these two epitopes is with higher intensity.

We divide the HA1 domain of the H3N2 hemagglutinin into six regions, namely epitopes A to E, and positions not in any of the epitopes. These regions show significantly distinct patterns of evolution. In each seasons from 1993–1994 to 2009–2010, we averaged selection and diversity in each epitope and the positions not in any of the epitopes. The fraction of positions j under selection defined by $S_{i,j} > S_i^{\text{thres}}$ was also calculated. The averages for 17 seasons are listed in Table 5.3. It is evident that the

Table 5.3 : Annual selection, fraction of positions under selection, and diversity in epitopes A to E, positions not in any of the epitopes, and the whole HA1 sequence.

Region	Selection	Fraction of positions under selection	Diversity
Epitope A	0.187	0.152	0.077
Epitope B	0.157	0.134	0.062
Epitope C	0.077	0.087	0.048
Epitope D	0.100	0.072	0.037
Epitope E	0.111	0.094	0.049
Out of epitopes	0.021	0.019	0.013
The whole sequence	0.060	0.051	0.028

values in Table 5.3 vary across the epitopes. The selection and diversity in epitopes A and B are greater than those in epitopes C, D, and E for each of selection, fraction of positions under selection, and diversity. The fraction of positions under selection is significantly greater than those in epitopes C, D, and E ($p < 0.038$, using Wilcoxon signed rank test). The values in epitopes C, D, and E are significantly greater than those not in any of the epitopes ($p < 0.0019$ for selection, $p < 0.0011$ for fraction of positions under selection, and $p < 6.0 \times 10^{-4}$ for diversity, using Wilcoxon signed rank test). Consequently, epitopes A and B display the highest level of selection and diversity.

5.4 Discussion

The Shannon entropy and the relative entropy are here introduced to quantify the diversity and the selection pressure of the evolving H3N2 virus. The foundation of the entropy calculation is the assumption that the virus sequences used in the entropy calculation are from a random unbiased sampling of the virus circulating in the human population. However, sampling density of the H3N2 virus varies in different geographic regions, hence creating a sampling bias.

We have addressed this issue at the continent level by analyzing data from different regions separately. We now additionally study the effect of sampling bias within one country. For example, among the H3 hemagglutinin sequences labeled with month of collection in the NCBI Influenza Virus Resource Database, the New York state sequences account for about one third of the USA sequences. In the contrast, New York state has only about 6.5% of the USA population. We chose eight seasons from 2001–2002 to 2008–2009 with abundant USA sequences collected in and out of New York state during this period of time available in the database. Using the procedure in the Materials and Methods section, we calculated the histogram of 20 amino acids in each amino acid position in each season for the New York state sequences, and that for the non-New York state sequences. Each of the $329 \times 8 = 2632$ pairs of histograms in and out of New York state were compared using the χ^2 test for homogeneity. The p values of 2581 pairs are greater than 0.05. That is, 98.1% of the pairs are not significantly different. The high sampling density in New York state does not affect the histograms of 20 amino acids in each position, implying that the sampling bias of the H3N2 virus is uncorrelated to the amino acid usage patterns.

By applying Shannon entropy and relative entropy to the aligned hemagglutinin sequences labeled with month of collection, we obtain the evolution and migration

pattern of the H3N2 virus in the Results section. First, Shannon entropy and relative entropy quantify diversity of and selection pressure over the virus, relatively. Relative entropy from the current season i to the next season $i+1$ linearly increased with the Shannon entropy in the current season i. See Figure 5.2 and 5.3. Second, relative entropy quantifies the similarity of two groups of virus and implies the migration path of the H3N2 virus. See Table 5.1. In the following text we compare our methods and results to the literature.

The relative entropy reveals the H3N2 migration pattern. Previous studies applied phylogenetic methods in an attempt to locate the epicenter of the H3N2 epidemic in each season. [32] studied the dynamic of influenza sequence diversity in the temperate regions in both hemispheres to imply that the H3N2 virus originates in the tropics and migrates to the temperate regions in both hemispheres. [43] obtained the antigenic and genetic evolution rate in each region and distances of the H3N2 strains to the trunk of the phylogenic tree. This information indicated East and Southeast Asia as the epicenter, from which the H3N2 virus spreads to North America, Europe, and Oceania in each season [43]. Recently, [146] suggested the center of the H3N2 migration network being China, Southeast Asia, and the USA by estimating the migration rate between different regions in the world. Here the relative entropy, the gain of sequence information, is used as a novel measure of the sequence similarity. The H3N2 migration path is the directed graph in which each path has the minimum relative entropy, or the maximum sequence similarity. These studies reach a consensus that South China is located in the epicenter of influenza epidemics. Here, we additionally identify a novel migration path from the USA to Europe and show that virus evolutionary rate is higher in the epicenter than in the migration paths.

Previous studies have identified positions that have led to the immune escape of in-

of the dN/dS ratio method. The second category operates at the amino acid level. [4] identified the positions with amino acid switch occurring in history to be under selection. Our entropy method recognizes the positions with relative entropy higher than the threshold, S_i^{thres}, in greater than two seasons. A large dN/dS of a codon does not necessarily mean an amino acid switch in the same position because the amino acid substitution could be unfixed. Methods at the amino acid level, such as the amino acid switch [4] and our entropy method, can identify positions with low dN/dS to be under selection because these methods do not consider nucleotide substitutions. Positive selection does not necessarily lead to a fixed amino acid switch, and in this case the entropy method can still detect positive selection. Unlike the amino acid switch method [4], the entropy method applied in this study is able to detect unfixed amino acid substitutions arising from selection. Our entropy method releases the requirement of fixed amino acid substitution in [4] but adds one requirement: the positions under selection need to present large relative entropy in greater than two seasons. Consequently, these methods identify slightly different sets of positions to be under selection.

5.5 Conclusion

We use Shannon entropy and relative entropy as two state variables of H3N2 evolution. The entropy method is able to predict H3N2 evolution and migration in the next season. First, we show that the rate of evolution increases with the virus diversity in the current season. The Shannon entropy data in one season strongly correlate with the relative entropy data from that season to the next season. If higher Shannon entropy of the virus is observed in one season, higher virus evolutionary rate is expected from this season to the next season. Second, the relative entropy values

between virus sequences from China, Japan, the USA, and Europe indicate that the H3N2 virus migration from China to Japan and the USA, and identify a novel migration path from the USA and Europe. The relative entropy values in and out of China, the epicenter, show that evolutionary rate is higher in China than in the migration paths. Moreover, the entropy method was demonstrated on two applications. First, selection pressure of the H3 hemagglutinin is mainly in 54 amino acid positions. Second, the top exposed part in the three-dimensional structure of HA trimer covered by epitopes A and B is under the highest level of selection. These results substantiate current thinking on H3N2 evolution, and show that the selection pressure is focused in a subset of amino acid positions in the epitopes, with epitopes A and B on the top of hemagglutinin being dominant and making the largest contribution to the H3N2 evolution. These predictions and applications show that the entropy method is not only predictive but also descriptive.

5.6 Appendix: Monte Carlo Simulation of the Patterns of Selection and Diversity

We introduce a Monte Carlo model aiming to regenerate the patterns of selection and diversity shown in Figure 4 in the main text. In this model, the sequence of the HA1 domain contains a dominant epitope bound by the antibody and possessing a high evolutionary rate $\mu_1 = 0.12$ amino acid substitution/site/season, with the other amino acid positions with a low evolutionary rate $\mu_2 = 0.0034$ amino acid substitution/site/season [42]. In each season, the numbers of positions in and out of the dominant epitope were defined as L_1 and L_2, respectively. An ensemble of 1000 HA1 sequences was created with identical amino acid identity in each of the 329 positions, and was simulated

from the 1969–1970 season to the 2009–2010 season. The historical dominant epitopes of H3 hemagglutinin were epitopes A and B, each of which was dominant for about seven seasons [2]. Therefore, the dominant epitope in the model was initialized as epitope A, and shifted between epitopes A and B every seven seasons. In each season, the numbers of amino acid substitutions in and out of the dominant epitope were randomly determined from two Poisson distributions with mean values $\lambda_1 = \mu_1 L_1$ and $\lambda_2 = \mu_2 L_2$, respectively. The positions of amino acid substitutions and the new amino acid identities were then randomly assigned. This process was repeated for all the 1000 sequences in each season. Following this procedure, we calculated for each position j the average relative entropy $\bar{S}_j$, the number of seasons under selection N_j, and the average diversity $\bar{D}_j$, using the simulated sequences between the 1993–1994 season and the 2009–2010 season. We also present the histograms of $\bar{S}_j$, N_j, and $\bar{D}_j$. The results of these calculations are shown in Figure 5.5.

The general picture depicted by this Monte Carlo simulation model reflects natural influenza evolution. The similarity between the results from the historical sequences in Figure 4 and those from the Monte Carlo simulation model in Figure 5.5 suggests that the Monte Carlo simulation model captures a major part of the picture of influenza evolution. The Monte Carlo simulation model also suggests that the binding of antibody and the increased substitution rate in the dominant epitope are the significant features of influenza evolution.

The annual evolution in the dominant epitope bound by the antibody decreases the affinity between virus and antibody and enables the virus to escape the immune memory of the virus circulating in the previous seasons. The evolutionary rate in the dominant epitope is $\mu_1 = 0.12$ amino acid substitution/site/season [42]. Both the historical sequences of the HA1 domain [4] and the Monte Carlo simulation model

suggest that substitutions of amino acid identities have occurred randomly across the dominant epitope. The relative frequency $f(k,i,j)$ of each amino acid k in each position j in each season i shows that the amino acid substitutions were in random positions in each season and displayed few visible correlation between two positions [4]. The Monte Carlo simulation model randomly selects the positions in which the amino acid is mutated, and generates similar patterns of selection and diversity to the historical data.

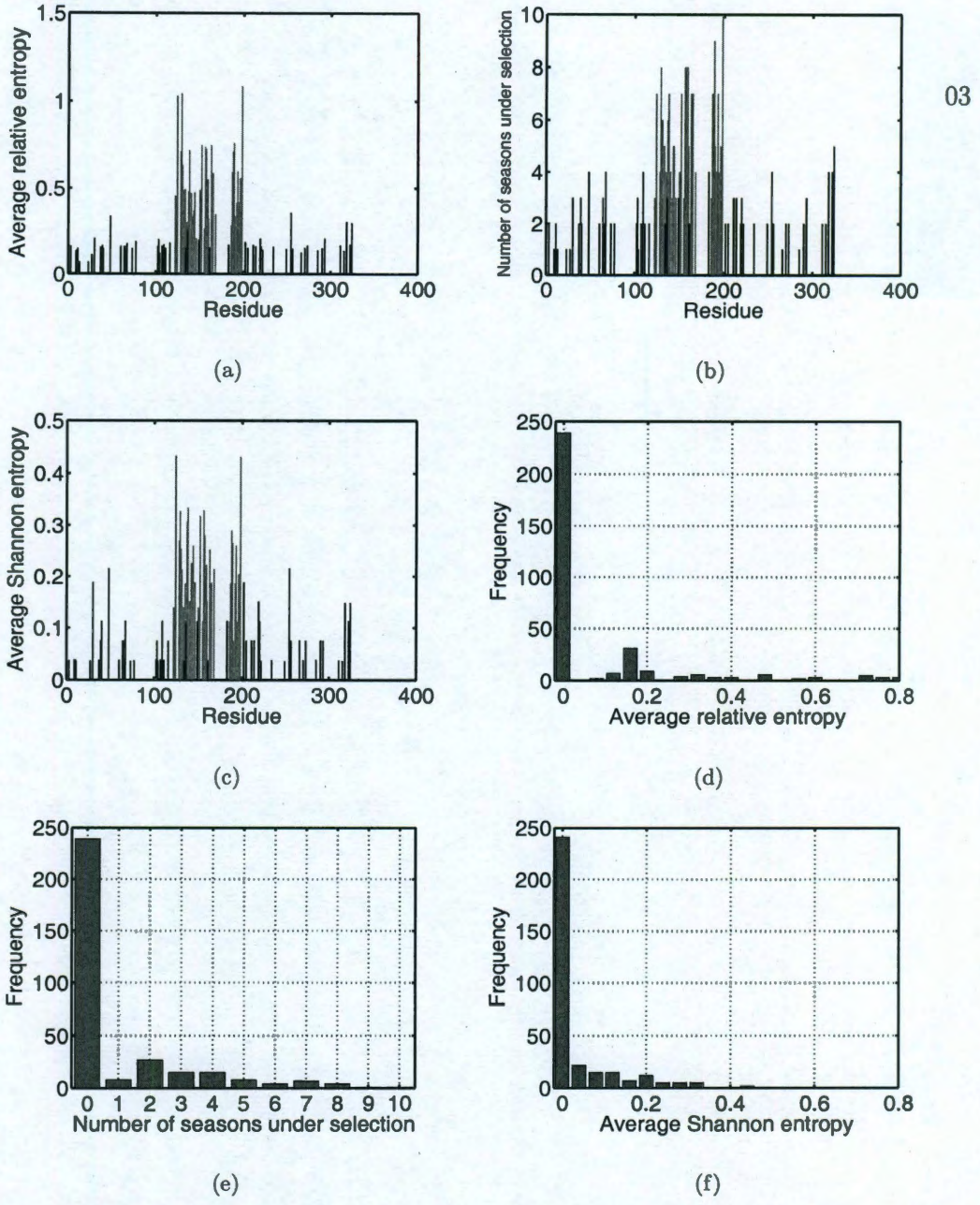


Figure 5.5 : The results of the Monte Carlo simulation model containing epitopes A (blue) and B (red), and all the other positions. The model was simulated for 41 seasons. (a) Average selection in each position quantified by relative entropy calculated by $\bar{S}_j = \sum_{i=1}^{17} S_{i,j}/17$ in the last 17 seasons. The colors represent positions in epitopes A and B and positions outside the epitopes. (b) Number of seasons for each position when the relative entropy was greater than the threshold S_i^{thres}, i.e. the position was under selection. (c) Average diversity in each position quantified by Shannon entropy in the 17 seasons, calculated by $\bar{D}_j = \sum_{i=1}^{17} D_{i,j}/17$. (d) Distribution of the average selection in each position displayed in (a). (e) Distribution of the numbers of seasons under selection displayed in (b). (f) Distribution of the average diversity in each position shown in (c).

Chapter 6

Selective Pressure to Increase Charge in Immunodominant Epitopes of the H3 Hemagglutinin Influenza Protein

The evolutionary speed and the consequent immune escape of H3N2 influenza A virus make it an interesting evolutionary system. Charged amino acid residues are often significant contributors to the free energy of binding for protein–protein interactions, including antibody–antigen binding and ligand–receptor binding. We used Markov chain theory and maximum likelihood estimation to model the evolution of the number of charged amino acids on the dominant epitope in the hemagglutinin protein of circulating H3N2 virus strains. The number of charged amino acids increased in the dominant epitope B of the H3N2 virus since introduction in humans in 1968. When epitope A became dominant in 1989, the number of charged amino acids increased in epitope A and decreased in epitope B. Interestingly, the number of charged residues in the dominant epitope of the dominant circulating strain is never fewer than that in the vaccine strain. We propose these results indicate selective pressure for charged amino acids that increase the affinity of the virus epitope for water and decrease the affinity for host antibodies. The standard PAM model of generic protein evolution is unable to capture these trends. The reduced alphabet Markov model (RAMM) model we introduce captures the increased selective pressure for charged amino acids in the dominant epitope of hemagglutinin of H3N2 influenza ($R^2 > 0.98$ between 1968 and 1988). The RAMM model calibrated to historical H3N2 influenza virus evolution in

humans fit well to the H3N2/Wyoming virus evolution data from Guinea pig animal model studies.

6.1 Introduction

Influenza A virus causes annual global epidemics resulting in severe morbidity and mortality. The dominant circulating virus today is the H3N2 virus, which emerged in 1968 and is defined by two kinds of surface glycoproteins: H3 hemagglutinin and N2 neuraminidase. It is currently believed that hemagglutinin is relevant to virus attachment and entry into the cell, while neuraminidase facilitates virus release [20]. Hemagglutinin also plays a central role in the process of immuno escape, in which the antibodies mainly attack five epitopes, denoted as epitopes A–E, on the surface of the hemagglutinin protein [59, 28]. Because of antigenic changes through time, influenza vaccines are redesigned each year to provide improved protection against evolved circulating strains. The efficacy of the annual vaccine is variable due to the escape mutation of the influenza virus [29], especially mutation at the five epitopes on the hemagglutinin [2].

By analyzing the results of over 50 epidemiological studies of H3N2 influenza during the period 1968 – 2004, [2, 60, 31, 33] showed that the escape mutation of influenza A virus can be measured by p_{epitope}, the proportion of mutated amino acids in the dominant epitope of hemagglutinin, where the dominant epitope is defined as the epitope with the largest such proportion among the five epitopes. Compared with p_{sequence}, the proportion of mutated amino acids in the whole sequence of hemagglutinin, and the ferret antisera assays, p_{epitope} between vaccine strains and dominant circulating strains in the same flu season correlated better with the vaccine efficacies in the northern hemisphere [2]. Therefore p_{epitope} is an appropriate measurement

for the antigenic distances between vaccine strains and dominant circulating strains. With the definition of the dominant epitope, the escape mutation at the dominant epitope induces the largest antigenic distance between vaccine strains and dominant strains, and endows the dominant epitope with the immunodominance.

In Gupta et al.'s model [2], which correlates well with vaccine efficacy in humans, every mutated amino acid is assigned the same weight. However, free energy calculations suggest that different amino acid substitutions have different contributions to the escape from the immune pressure. In general, the calculated differences in binding free energy $\Delta\Delta G = \Delta G_{\text{mutated}} - \Delta G_{\text{wildtype}}$ are different for different mutations, where $\Delta G_{\text{mutated}}$ and $\Delta G_{\text{wildtype}}$ denote binding free energy between two proteins one of which has and does not have a point mutation, respectively. For the experimentally measured difference in binding free energy $\Delta\Delta G$ between human growth hormone (hGH) and its first bound receptor (hGHbp), individual alanine substitutions of hydrophobic amino acids on the epitope of hGHbp induced the largest increase in $\Delta\Delta G$, followed by charged amino acids [8]. Nevertheless, we show that charged amino acids correlate more strongly with viral evolution (Table 6.3 and 6.4 in this chapter). Nakajima [9] found that the majority of escape mutations of H3 hemagglutinin of the strain A/Kamata/14/91 were the mutations that introduce charged amino acids, and that the frequency of selected mutations to charged residues was significantly higher than that expected by random chance. A related study by Smith [30] found a similar over-representation of mutations to charged amino acids in the evolution of influenza.

We here consider the effect of different physical properties on the escape from immune pressure. We focus on the charged amino acids in the epitopes. These amino acids are strongly hydrophilic, and they reduce the tendency of antibodies to bind to hemagglutinin. Charged amino acids play a critical role in protein–protein interaction

by creating salt bridges and salt bridge networks. Charged amino acids introduce specificity in binding [10]. Amino acid substitutions involving charged residues in the vicinity of receptor-binding region of HA affect the binding affinity between HA and its receptor [11]. The evolution of charged amino acids, therefore, may provide useful information on viral escape from antibody pressure. A discussion of charge evolution in proteins has been given by Leunissen [12], who observed a large variance in the evolutionary trends among different protein families. Here we use stochastic methods to model the evolution of charged amino acids on the epitopes of H3 hemagglutinin strains collected from humans since 1968.

The metaanalysis of 50 epidemiological human vaccine efficacy studies shows that the single dominant epitope is the critical region that determines the epidemiological vaccine efficacy [2]. There are five non-overlapping epitopes on the surface of H3 HA molecule, namely epitope A–E, to which different sets of antibodies bind. In each epitope, the p value is defined as the fraction of mutated amino acids [2]. The dominant epitope is defined as the epitope with the greatest p value. The greatest p value is p_{epitope}. Epidemiological data on the vaccine efficacies in 18 previous flu seasons when H3N2 subtype was dominant were collected from approximately 50 studies [2]. The identities of the vaccine strains and dominant circulating strains were also obtained to calculate p_{epitope}. H3N2 vaccine efficacy correlates with p_{epitope} with $R^2 = 0.81$. This strong correlation shows that p_{epitope} defined by the single dominant epitope is a quantitative definition of antigenic distance. Importantly, the p_{epitope} calculated from the dominant epitope correlated better with vaccine efficacy than did antigenic distance including all HA amino acids [2].

The results of [61] show that subdominant epitopes are not the critical regions for vaccine efficacy. In an effort to improve the definition of antigenic distance, four

modifications of the definition of antigenic distance were tested for their ability to improve the correlation with vaccine efficacy: 1) incorporating p values from subdominant epitopes, 2) distinguishing conservative and non-conservative amino acids mutations, 3) mutations in amino acids adjacent to the epitopes, and 4) the calculation of mutations in neuraminidase [61]. These four modifications of the definition of antigenic distance, including use of subdominant epitope p values, all failed to substantially improve the correlation with vaccine efficacy data in the years 1971–2004. These results motivate our focus on the dominant epitiope in the present analysis.

The reduced alphabet Markov model (RAMM) described in this chapter is an amino acid substitution model. It is built from the H3 hemagglutinin strains circulating in 1972 – 1987 when epitope B was the dominant epitope. This time span is shortly after the emergence of H3N2 virus in 1968, and this newly emerged virus subtype needed some time to adapt to host immune system, because emergence of new subtypes such as H2N2 in 1957 and H3N2 in 1968 went with Asian flu and Hong Kong flu outbreaks, indicating that subtypes like H3N2 were more virulent in the beginning and less adaptive to human. Further, the phylogenetic tree of H3N2 also shows that H3N2 evolved faster at the very beginning than in the later stage [30]. So the pattern of evolution illustrates the escape mutation of the virus before substantial adaptation to host immune system was developed. Because the sequence database has been derived from patient samples and categorized by antigenic strain and date of collection, the human data do not necessarily reflect fixed variants, but, rather, snapshots along an evolutionary continuum.

Mutations of amino acids in different positions in the epitope are viewed as independent and identical Markov chains, whose parameters are the transition matrix $\mathbf{P}$ or the instantaneous rate matrix $\mathbf{Q}$. Markov models of protein evolution include

the point accepted mutation (PAM) model [62] and the block substitution matrix (BLOSUM) model [63]. These models are derived by counting mutations in aligned amino acid sequences, and this approach provides the transition matrices $\mathbf{P}(t)$ of a Markov chain in a period of evolutionary time t. [64] introduced the maximum likelihood method to estimate the elements of the transition matrix, and maximum likelihood was also employed to estimate the evolutionary time t when fixing the transition matrix $\mathbf{P}(t)$ [65]. The instantaneous rate matrix $\mathbf{Q}$ was calculated from the Laplace transform of $\mathbf{P}(t)$ [65, 66]. For a review of applications to 2000, see [67]. Some recent studies estimated the effect of possible multiple mutations at the same position within evolutionary time t [68, 69]. The instantaneous rate matrix has been estimated from observed frequencies of 20 amino acids [70].

The transition matrices $\mathbf{P}$ and the instantaneous rate matrices $\mathbf{Q}$ of most previous models are 20×20 matrices trained by analyzing databases with numerous alleles in many taxa. In the present case, the training data are limited, which can cause overfitting and introduce large errors to the fit model. One way to circumvent this difficulty is to decrease the number of parameters: 20 amino acids may be classified into several groups with similar biophysical properties. Here we grouped 20 amino acids as 5 charged amino acids and 15 uncharged amino acids.

To test whether the mathematical models can be applied to data outside of the existing human data, we investigated their relevance to animal models of influenza infection. Guinea pigs have been shown to be infected with unadapted H3N2 strains. In most cases, the infection is limited to the upper respiratory tract, causes little apparent morbidity, and can be spread to cage mates via aerosol [71]. As part of our analysis of the p_{epitope} calculations, we developed additional evolutionary data derived through analysis of progeny viruses in animal model systems. Interestingly, the pat-

tern of variations in the Guinea pig infection correlated well with the predictions of the model and with the human evolutionary data.

6.2 Materials and Methods

We defined the charged amino acids as {Asp, Glu, Arg, Lys, His}. Historical sequence data indicated that the number of charged residues increased continually on the dominant epitope of the hemagglutinin of the circulating H3N2 strains in Figure 6.1, except for the data in the year 1996 – 1997 which was presumably a continuation of the decreasing trend in the charge of epitope B since 1993. The vaccine strains in Figure 6.2 also had increasing numbers of charged amino acids in a period of time without epitope shift. The charge in the dominant epitope sometimes stopped increasing, which means the existing of the steady state for the charge and will be discussed later. Moreover, since the observed dominant epitope in a given year was either epitope A or epitope B, we defined the subdominant epitope as the epitope in the set {epitope A, epitope B} other than the dominant one. Figure 6.1 and Figure 6.2 also shows that the number of charged residues decreased in the subdominant epitope in both the circulating strains and the vaccine strains. For vaccine strains and circulating strains collected in 19 years, vaccine efficacies in each year had a strong correlation ($R^2 = 0.81$) with the p_{epitope} in that year [2], but R^2 decreased quickly if we use the weighted sum of p_{epitope} and p_{epitope} in previous years (Pan and Deem, unpublished data). Therefore we treated the immune escape of the virus as a Markov process [149]. Additionally, since there is limited information on the correlation between different positions on a given epitope, we assumed that the mutations at different positions on the epitope followed identical and independent Markov chains.

The epitope A and epitope B of the circulating strains and the strains deposited

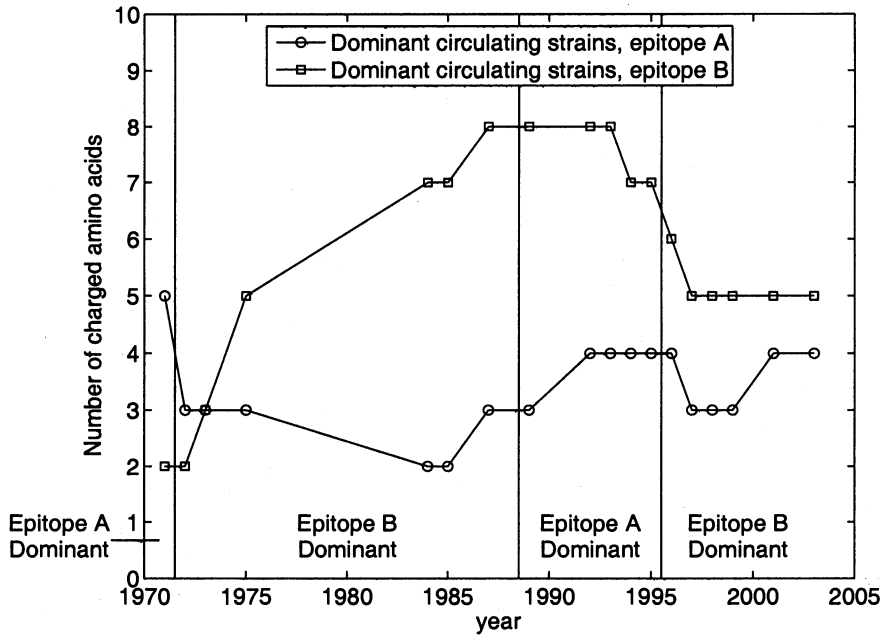


Figure 6.1 : Number of charged amino acids for each year on the epitope A and epitope B of the dominant circulating strains from 1971 to 2003.

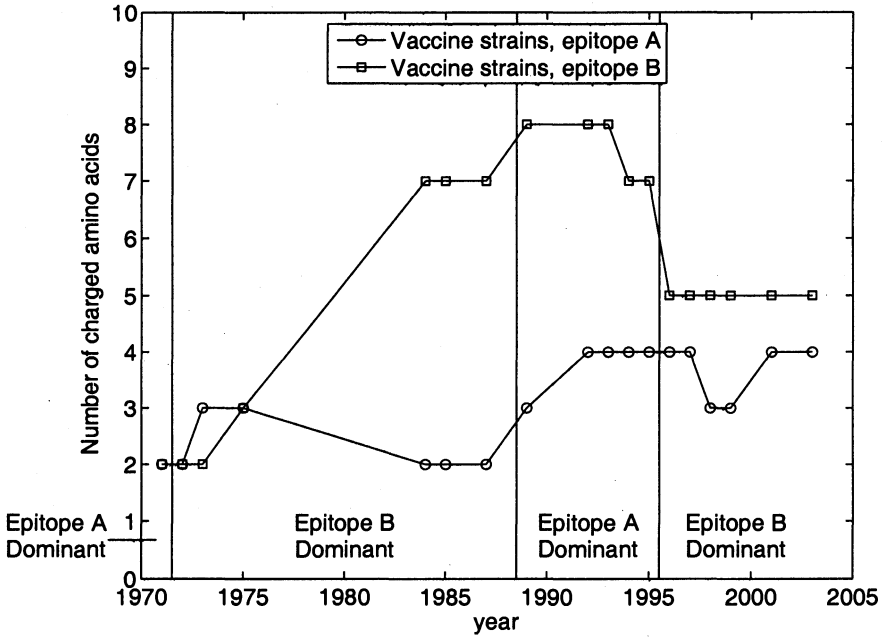


Figure 6.2 : Number of charged amino acids for each year on the epitope A and epitope B of the vaccine strains from 1971 to 2003. There were four consecutive intervals where epitope A or epitope B was dominant.

in the GenBank database were collected in 18 years between 1968 and 2003 and listed in Table 6.1, where the numbers of charged amino acids are presented as number for the dominant circulating strain and the average number for the circulating strains in GenBank database in the same year [2]. These data were also plotted (Figure 6.1 and Figure 6.3). The dominant epitope was epitope A in the 1971 strain, epitope B in the {1972, 1973, 1975, 1984, 1985, 1987} strains, epitope A in the {1989, 1992, 1993, 1994, 1995} strains, and epitope B in the {1996, 1997, 1998, 1999, 2001, 2003} strains. All the strains except the 1971 strain fell into three multi-year time intervals defined by unchanged dominant epitopes. The numbers of charged amino acids on the dominant epitopes of each strain were counted and utilized. The total number of charged amino acid on a certain epitope, N_c, was modeled by a binomial distribution with probability

$$\Pr\{N_c = n\} = \binom{N}{n} P_c^n(t \mid \epsilon, \delta_1)\left[1 - P_c(t \mid \epsilon, \delta_1)\right]^{N-n} \tag{6.1}$$

with the mean number of amino acids in the epitope equal to $NP_c(t \mid \epsilon, \delta_1)$, where N was the total number of amino acids on the epitope.

Table 6.1 : Epitope A and B of dominant circulating strains. Two numbers of charged amino acids in each epitope in each year are presented; the first one is calculated from the dominant circulating strain, and the second one is the simple arithmetic average of all strains collected in that year and deposited in GenBank.

Year	Dominant circulating strain	Dominant epitope	Number of charged amino acids in epitope A	Sequence of epitope A of the dominant circulating strain	Number of charged amino acids in epitope B	Sequence of Epitope B of the dominant circulating strain
1971–1972	HongKong/1/68 (AF201874)	A	5/2.22	TGTVTQDGNAKGPGKRRNM	2/2.78	TGTKSGSTVNSTNQETSLVQA
1972–1973	England/42/72 (AF201875)	B	3/2.95	NGTVTQNGNAKGPDSSRNM	2/3.00	TGYKSECTVNSTNQVTSLVQA
1973–1974	PortChalmers/1/73 (AF092062)	B	3/3.00	NGTVTQNGNAKGPDSGRNM	3/3.00	TGYKSGSAVNSTDQETNLVQA
1975–1976	Victoria/3/75 (ISDNVIC75)	B	3/3.18	NGNVTQNGSAKGPDNGRNM	5/4.18	TGYKLGSTVNSTDKETDLVQA
1984–1985	Mississippi/1/85 (AF008893)	B	2/2.00	NGNVTQSGYAKGSVNSRNM	7/6.88	TGYKSESKANSTDKETNLVRA
1985–1986	Mississippi/1/85 (AF008893)	B	2/2.13	NGNVTQSGYAKGSVNSRNM	7/7.03	TGYKSESKANSTDKETNLVRA
1987–1988	Shanghai/11/87 (AF008886)	B	3/3.20	NDNVTQSGYAKGSVNSRNM	8/8.20	TGHESEYKANSTDRETNLVRA
1989–1990	England/427/88 (AF204238)	A	3/3.42	NDNVAQSGCAKGSVNSRNM	8/8.20	TGHESEYKANSTDRETNLVRA
1992–1993	Beijing/32/92 (AF008812)	A	4/4.64	NDNVAQDGYAKGSVNSRNM	8/7.79	TGHKSEYKANSTDRDTSLVRA
1993–1994	Beijing/32/92 (AF008812)	A	4/4.77	NDNVAQDGYAKGSVNSRNM	8/7.02	TGHKSEYKANSTDRDTSLVRA
1994–1995	Johannesburg/33/94 (AF008774)	A	4/4.71	NNNVAQDKYAKGSVNSRNM	7/6.94	TGHKLEYKANSTDSDTSLVRA
1995–1996	Johannesburg/33/94 (AF008774)	A	4/4.35	NNNVAQDKYAKGSVNSRNM	7/6.36	TGHKLEYKANSTDSDTSLVRA
1996–1997	Wuhan/359/95 (AF008722)	B	4/4.01	NGNVAQDTYAKGSVKSRNM	6/6.01	TGHKLEYKANSTDSDTSIVQA
1997–1998	Sydney/5/97 (AJ311466)	B	3/4.07	NSNVAQNTYAKSSIKSRNM	5/5.53	TGHQLKYKANSTDSDTSIAQA
1998–1999	Sydney/5/97 (AJ311466)	B	3/4.03	NSNVAQNTYAKSSIKSRNM	5/5.18	TGHQLKYKANSTDSDTSIAQA
1999–2000	Sydney/5/97 (AJ311466)	B	3/3.95	NSNVAQNTYAKSSIKSRNM	5/5.01	TGHQLKYKANSTDSDTSIAQA
2001–2002	Panama/2007/99 (ISDNCDA001)	B	4/4.70	NSNVAQNTSAKRSNKSRNM	5/5.02	TGHQLKYKANSTDSDISLAQA
2003–2004	Fujian/411/2002 (ISDN38157)	B	4/4.53	NSNVTQNTSAKRSNKSRNM	5/4.98	TGTHLKYKANGTDSDISLAQA

6.2.1 Discrete–Time Markov Chain

We first applied a discrete–time Markov chain using one year as the time unit because the available data are the strains deposited in databases with one–year time resolution. The probability distribution was defined as $\pi(t) = (P_c(t), P_u(t)) = (P_c(t), 1 - P_c(t))$ with the initial value $(P_{c0}, 1 - P_{c0})$, where P_c and P_u were the probabilities that a charged amino acid and an uncharged amino acid existed in a given position on the dominant epitope. Following this definition, the transition matrix is a 2–by–2 matrix, taking parameters that describes the evolutionary process. We chose the mutation probability for each position, ϵ, and the bias for mutation to charged amino acids from charged or uncharged amino acids, δ_1 and δ_2, respectively, to characterize the virus evolution model discussed in this chapter. Thus the transition matrix is

$$\mathbf{P} = (1-\epsilon)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \epsilon \begin{pmatrix} \alpha_{cc} & \alpha_{cu} \\ \alpha_{uc} & \alpha_{uu} \end{pmatrix} \tag{6.2}$$

$$\text{with} \quad \alpha_{cc} = \frac{5}{20} + \delta_1, \alpha_{cu} = \frac{15}{20} - \delta_1, \alpha_{uc} = \frac{5}{20} + \delta_2, \alpha_{uu} = \frac{15}{20} - \delta_2$$

where ϵ is the amino acid substitution rate (the new amino acid could be identical to the original amino acid); and $\alpha_{cc}, \alpha_{cu}, \alpha_{uc}$, and α_{uu} are the conditional probabilities that a charged residue mutated to a charged residue, a charged residue mutated to an uncharged residue, an uncharged residue mutated to a charged residue, and an uncharged residue mutated to an uncharged residue, respectively, if the substitution occurred. The parameters δ_1 and δ_2 were the bias for the probability that a charged amino acid mutated to a charged amino acid, an uncharged amino acid mutated to a charged amino acid, respectively. Under the further assumption that $\delta_1 = \delta_2 = \delta$,

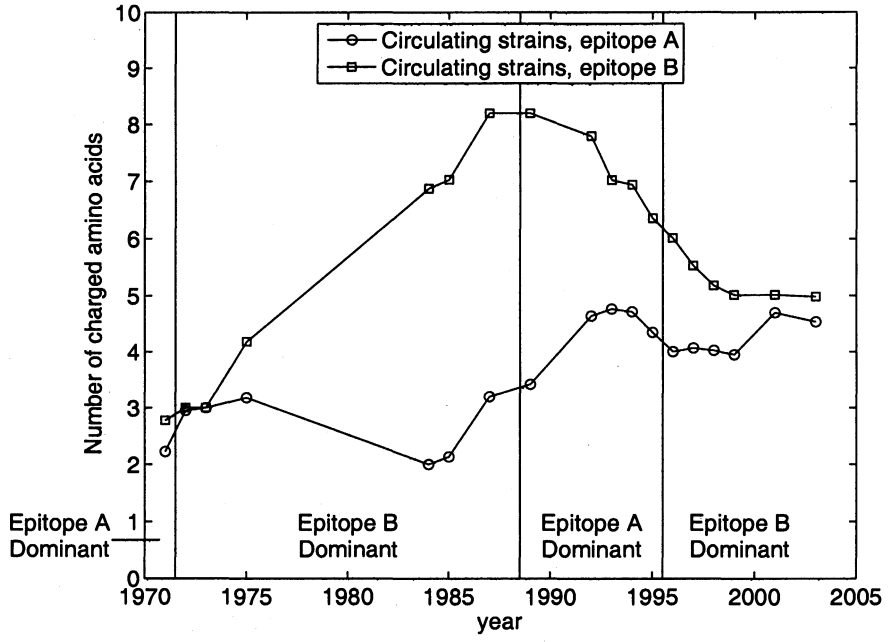


Figure 6.3 : Average number of charged amino acids for each year on the epitope A and epitope B of the strains deposited in the GenBank database from 1971 to 2003. The curves in this figure were smoother than those in Figure 6.1 due to the averaging over database strains. The difference in the numbers of charged amino acids between this figure and Figure 6.1 is smaller than one for most years.

the probability distribution in year t is

$$\begin{aligned}\left(P_{\mathrm{c}}\left(t \mid \epsilon, \delta\right), P_{\mathrm{u}}\left(t \mid \epsilon, \delta\right)\right) &= \left(P_{\mathrm{c0}}, 1-P_{\mathrm{c0}}\right)\mathbf{P}^{t} \\ &= \left(P_{\mathrm{c0}}, 1-P_{\mathrm{c0}}\right)\mathbf{Q}^{-1}\mathbf{D}^{t}\mathbf{Q} \\ &= \left(\tfrac{1}{4}\left(1+4\delta-(1-\epsilon)^{t}\left(1+4\delta-4P_{\mathrm{c0}}\right)\right), 1-P_{\mathrm{c}}\left(t \mid \epsilon, \delta\right)\right) \end{aligned} \tag{6.3}$$

where the matrix $\mathbf{Q}$ was invertible, the matrix $\mathbf{D}$ was diagonal.

6.2.2 Continuous–Time Markov Chain

Besides the discrete–time Markov chain, the continuous–time Markov chain was also commonly employed in several bio-related fields including the estimation of residue mutation rates [150]. In the continuous–time Markov chain, the transition rate matrix $\mathbf{Q}$ is first defined describing the evolution of probability distribution in a infinitesimal time interval, with $\mathbf{Q1} = \mathbf{0}$ where $\mathbf{1}$ is a column vector with all the elements equal to unity. The transition matrix is

$$\mathbf{P}\left(t\right) = \begin{pmatrix} P_{\mathrm{cc}}\left(t\right) & P_{\mathrm{cu}}\left(t\right) \\ P_{\mathrm{uc}}\left(t\right) & P_{\mathrm{uu}}\left(t\right) \end{pmatrix} = \exp\left(\mathbf{Q}t\right) \tag{6.4}$$

with the probability distribution at year t was again defined as $\boldsymbol{\pi}\left(t\right) = \left(P_{\mathrm{c}}\left(t\right), P_{\mathrm{u}}\left(t\right)\right) = \boldsymbol{\pi}\left(0\right)\mathbf{P}\left(t\right)$ with the initial probability distribution $\boldsymbol{\pi}\left(0\right) = \left(P_{\mathrm{c0}}, 1-P_{\mathrm{c0}}\right)$. And if the Markov chain contains finite states, the Kolmogorov backward equation (KBE)

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{P}\left(t\right) = \mathbf{Q}\mathbf{P}\left(t\right) \qquad \text{(KBE)} \tag{6.5}$$

holds.

For the evolution process of charged amino acid, the $\mathbf{Q}$ matrix was defined as

$$\mathbf{Q} = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix} \qquad \lambda_1, \lambda_2 \geq 0. \tag{6.6}$$

The solution for the transition matrix $\exp(\mathbf{Q}t)$ was

$$\begin{aligned}
P_{\mathrm{cc}}(t) &= \frac{\lambda_2}{\lambda_1+\lambda_2} + \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)t} \\
P_{\mathrm{cu}}(t) &= \frac{\lambda_1}{\lambda_1+\lambda_2} - \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)t} \\
P_{\mathrm{uc}}(t) &= \frac{\lambda_2}{\lambda_1+\lambda_2} - \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)t} \\
P_{\mathrm{uu}}(t) &= \frac{\lambda_1}{\lambda_1+\lambda_2} + \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)t}.
\end{aligned}$$

The term P_{c} is given by

$$P_{\mathrm{c}}(t \mid \lambda_1, \lambda_2, P_{\mathrm{c0}}) = \frac{\lambda_2}{\lambda_1+\lambda_2} + \left(P_{\mathrm{c0}} - \frac{\lambda_2}{\lambda_1+\lambda_2}\right) e^{-(\lambda_1+\lambda_2)t}. \tag{6.7}$$

6.2.3 Maximum Likelihood Estimation

The maximum likelihood estimation method optimizes the parameters in a given parametric form by maximizing the log-likelihood function with the definition $l(\boldsymbol{\theta} \mid \mathbf{x}) = \ln P(\mathbf{x} \mid \boldsymbol{\theta})$. The observed data $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ are usually independent, therefore the maximum likelihood function is $l(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{k=1}^{m} \ln P(\mathbf{x}_k \mid \boldsymbol{\theta})$. In our model, the data $\mathbf{x}_k = (n, t)$ in different years were assumed to be independent, where n was the number of charged amino acids, t was the time in years, and the parameter set was $\boldsymbol{\theta} = (\epsilon, \delta, P_{\mathrm{c0}})$. From the relationship between the total number of charged amino acids and the probability distribution at each position (6.1), the maximum likelihood equation was written as

$$\begin{aligned}
l(\epsilon, \delta, P_{\mathrm{c0}} \mid \mathbf{x}) &= \sum_{k=1}^{m} \ln P(\mathbf{x}_k \mid \boldsymbol{\theta}) \\
&= \sum_{k=1}^{m} \ln \binom{N}{n_k} + \sum_{k=1}^{m} n \ln P_{\mathrm{c}}(t_k \mid \epsilon, \delta, P_{\mathrm{c0}}) \\
&\quad + \sum_{k=1}^{m} (N-n) \ln\left[1 - P_{\mathrm{c}}(t_k \mid \epsilon, \delta, P_{\mathrm{c0}})\right]
\end{aligned} \tag{6.8}$$

with the expression of P_c in (6.3).

The maximum likelihood function was maximized by solving the following equations

$$\begin{cases} \frac{\partial}{\partial \epsilon} l\left(\epsilon, \delta, P_{c0} \mid \mathbf{x}\right) &= 0 \\ \frac{\partial}{\partial \delta} l\left(\epsilon, \delta, P_{c0} \mid \mathbf{x}\right) &= 0 \\ \frac{\partial}{\partial P_0} l\left(\epsilon, \delta, P_{c0} \mid \mathbf{x}\right) &= 0 \end{cases} . \tag{6.9}$$

The numerical solution was identical to that of the least square fit to the first two significant figures. We note that solutions of (6.9) were not ill-conditioned while the nonlinear fitting algorithm for (6.3) led to ill-conditioned Jacobians.

6.2.4 Guinea Pig Animal Model

The Guinea pig model of human influenza infection was adopted to model the process of infection and reinfection [71]. A sample of A/Wyoming/2003 (H3N2) was obtained from the Centers for Disease Control and Prevention (CDC). Sequence analysis of multiple plaque isolates from the CDC virus demonstrated that the sample contained a mixture of viruses with several hemagglutinin (HA) sequences, all in close agreement with the A/Wyoming/03/2003 sequence reposited in GenBank (accession number AAT08000). An isolate (WyB4) representing the dominant HA sequence contained within the CDC virus mixture was purified and propagated.

In the first infection experiment, four immunologically naïve Guinea pigs were inoculated with the CDC virus mixture to model virus evolution in the absence of robust immunity. Three days following inoculation, progeny virus were purified from nasal washes for sequence analysis of the HA genes. After a recovery period of 28 days, two of the animals were reinfected with the CDC virus mixture to model virus escape in the presence of increased immune pressure. Nasal washes were collected

after 3 days, progeny virus isolated, and HA gene sequences analyzed. In a second infection experiment, six naïve Guinea pigs were inoculated with the purified WyB4 isolate. Again, virus progeny were purified from nasal wash samples for HA sequence analysis. In a third infection experiment, Guinea pigs were inoculated with three immunizations of recombinant Wyoming HA protein having the same sequence as the WyB4 isolate. After the animals had mounted robust immune responses, they were challenged with either WyB4 or the CDC virus mixture. Progeny virus were isolated from nasal washes for HA gene sequence analyses.

6.3 Results

6.3.1 Charge Increases in the Dominant Epitope

The parameters obtained by the fitting algorithm for the first time interval (epitope B was dominant, from 1972 to 1987) were $\boldsymbol{\theta}^{\mathrm{T}} = (\epsilon, \delta, P_{\mathrm{c0}}) = (0.207, 0.113, 0.0937)$ with the square of the correlation coefficient between observations and the model $R^2 = 0.98$. The mathematical details are shown in Supplementary Material. Here the transition matrix was

$$\mathbf{P} = \begin{pmatrix} 0.868 & 0.132 \\ 0.075 & 0.925 \end{pmatrix}. \tag{6.10}$$

We plotted the observed data and the data predicted by the model with the parameters fixed by maximum likelihood estimation in Figure 6.4. We fit ϵ and δ to the data in the range 1972–1987 and used these values for all years. We fit the value of P_{c0} in the range 1972–1987 and found it equal to the data point at 1972. For other intervals we used the observed initial values.

The model parameters here yielded the average number of point mutations in epitope B $= 19N\epsilon/20 = 4.07$, and the observed numbers of point mutation in epitope

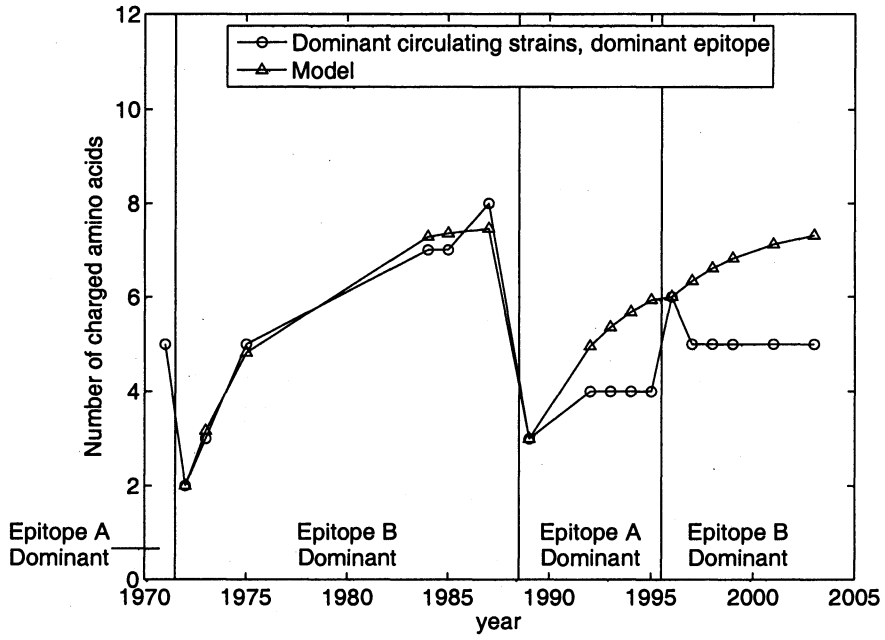


Figure 6.4 : Number of charged amino acids for each year on the dominant epitopes of the circulating strains. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 are plotted. Since the estimated P_{c0} was close to the observed number, we used the observed numbers of charged amino acids in the first year of the interval to calculate P_{c0}.

B during one year were: 1971–1972, 4; 1972–1973, 6; 1984–1985, 0. Those are all the available one-year time spans in the training data, and this model falls inside the standard error of the actual annual mutation rates on the dominant epitopes. That $\delta = 0.11$ showed that the conditional probability that a charged residue mutated to a charged residue was 0.36 while the probability that an uncharged residue mutated to an uncharged residue was 0.64. The observed number of charged amino acids on epitope B in 1972 was 2, and corresponding $P_{c0} = 2/21 \approx 0.095$, which agrees well with the fit value $P_{c0} = 0.094$.

We also obtained all the strains deposited in the GenBank database that were collected in the same year as a circulating strain existed. The numbers of charged amino acids $N_c = NP_c(t \mid \epsilon, \delta, P_{c0})$ were counted for each year as the mean value of the numbers of charged amino acid on the dominant epitope of the strains collected in that year. The standard deviations of the numbers of charged amino acid on the dominant epitope were also calculated for each year. Again, there is no assurance that any of these mutations can be considered end-point or fixed variants as the evolutionary process continued from year to year. By using these $P_c(t \mid \epsilon, \delta, P_{c0})$ to fit the parameters $\boldsymbol{\theta} = (\epsilon, \delta, P_{c0})$, ϵ approached to zero when the number of iterations of the fitting algorithm increased to infinity, while the product $\epsilon\delta$ approached to 0.016 compared with the product $\epsilon\delta = 0.023$ in the case with only circulating strains. Both the calculated and observed P_{c0} equal to 0.14. The model fit well with all the datapoints in this time span with $R^2 = 0.99$. We note that the Markov model is not able to reproduce convex and rising data, and this is the reason for $\epsilon \to 0$ for these data. In fact the data in Figure 6.5 are convex probably because the simple arithmetic averaging is only a rough approximation, and we expect concave and rising data, which the Markov model can represent, from a properly weighted population

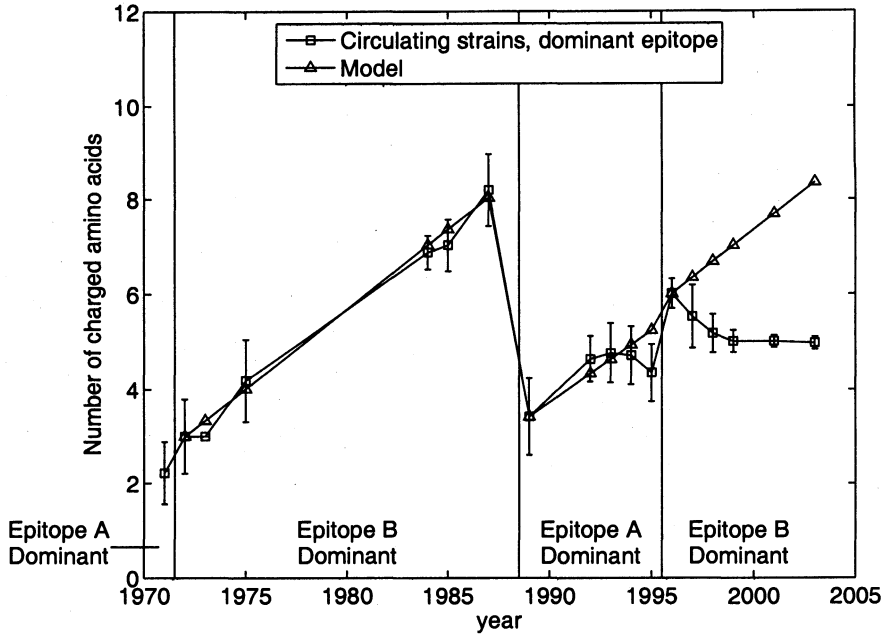


Figure 6.5 : Number of charged amino acids for each year on the dominant epitopes of the average of all the database strains in the same year. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 were plotted. We used the observed numbers of charged amino acids in the first year of the interval to calculate P_{c0}. Error bars for the circulating strains are one standard deviation calculated from all NCBI strains collected in that year.

average. The maximum likelihood method estimated the same parameter values to within two significant figures as the least square fitting method.

The continuous–time Markov chain possessed the similar parameter set $\boldsymbol{\theta} = (\lambda_1, \lambda_2, P_{c0})$ to the discrete–time Markov chain. For the circulating strains in 1972–1987, the nonlinear least-squares fitting yielded the parameters $\lambda_1 = 0.15, \lambda_2 = 0.084, P_{c0} = 0.094$ with $R^2 = 0.98$. For the database strains, there were $\lambda_1 = -0.031, \lambda_2 = 0.011, P_{c0} = 0.14$ with $R^2 = 0.99$. As before, simple arithmetic averaging produces convex and rising data, which the Markov model is not able to

represent, and which is not expected from a population average.

The RAMM model fit well the data in circulating strains and those in the GenBank database in the first time interval (1972–1987). This model had a larger discrepancy in the prediction versus both groups of data in the second (1989–1995) and third time interval (1996–2003).

6.3.2 Evolution of Amino Acids in Epitopes of Hemagglutinin is Non–universal: Comparison with PAM Matrix

To compare with the RAMM model, we also employed the conventional PAM matrix [62] to predict the evolution of charged amino acids. We first calculated the PAM1 matrix $\mathbf{M} = \{M_{ij}\}$ describing the evolution of one amino acid position during a time unit in which the probability that a point mutation occurred in this position was fixed by Dayhoff et al. [62] to 0.01 (the meaning of the suffix 1). Evolution during n time units was depicted by PAMn matrix with PAMn = PAM1^n = $\mathbf{M}^n = \{M'_{ij}\}$. The probability for a point mutation occur was then

$$P = \sum_{i=1}^{20} (1 - M'_{ii})\, \pi_i. \tag{6.11}$$

We let this probability equal to the annual mutation rate calculated by the RAMM model and verified by the historical data, $19\epsilon/20 = 0.196$, and so PAM22 should be selected to calculate the mutations during one year. The more frequently used PAM20 matrix, however, underpredicts the annual mutation rate and generates larger errors. While the PAM22 matrix reproduces the observed number of total mutations in the dominant epitope, as we shall see, it underpredicts the number of charged mutations per year. This means there is an additional selective force for charged amino acids in the epitopes of hemagglutinin relative to protein evolution in general. To calculate

the 2×2 transition matrix with the definition

$$\mathbf{P} = \begin{pmatrix} P_{cc} & P_{cu} \\ P_{uc} & P_{uu} \end{pmatrix}$$

there were

$$\begin{aligned} P_{cu} &= P\{\text{charged} \rightarrow \text{uncharged} \mid \text{charged}\} \\ &= \frac{\sum_{i=\text{charged}} \left(\sum_{j=\text{uncharged}} M'_{ij} \right) \pi_i}{\sum_{i=\text{charged}} \pi_i} \qquad (6.12) \\ P_{uc} &= \frac{\sum_{i=\text{uncharged}} \left(\sum_{j=\text{charged}} M'_{ij} \right) \pi_i}{\sum_{i=\text{uncharged}} \pi_i} \qquad (6.13) \\ P_{cc} &= 1 - P_{cu} \qquad (6.14) \\ P_{uu} &= 1 - P_{uc}. \qquad (6.15) \end{aligned}$$

yielding the solution of the transition matrix $\mathbf{P}$ based on PAM22 matrix

$$\mathbf{P} = \begin{pmatrix} 0.883 & 0.117 \\ 0.040 & 0.960 \end{pmatrix} \qquad (6.16)$$

which is slightly different from RAMM matrix (6.10). The evolution predicted by PAM22 matrix for the circulating strains and equally weighted database strains is plotted in Figure 6.6 and Figure 6.7, respectively.

6.3.3 Guinea Pig Animal Model Verifies the Increase in Charge

To buttress our analyses of the evolution of charged amino acids in historical virus sequences, we examined the frequency of charge changes in the HA sequences derived from the progeny virus using the Guinea pig animal infection model. The animals were divided into two groups, naïve and immune. The immune animals were either previously infected or immunized with HA protein. The animals were given either

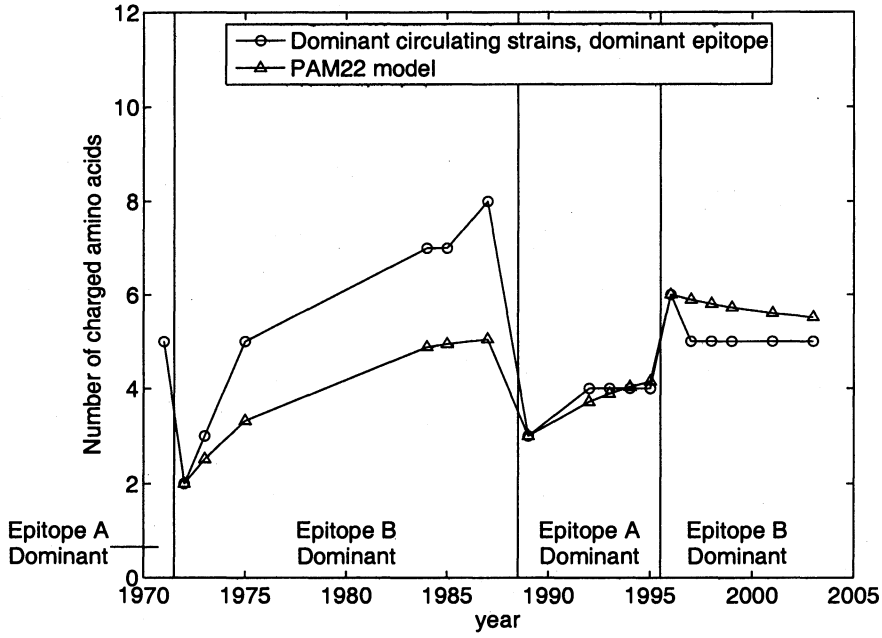


Figure 6.6 : Number of charged amino acids for each year on the dominant epitopes of the circulating strains. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 with PAM22 matrix were plotted. P_{c0} was fixed to the observed numbers of charged amino acids in the first year of the interval. The PAM22 matrix cannot reproduce the data.

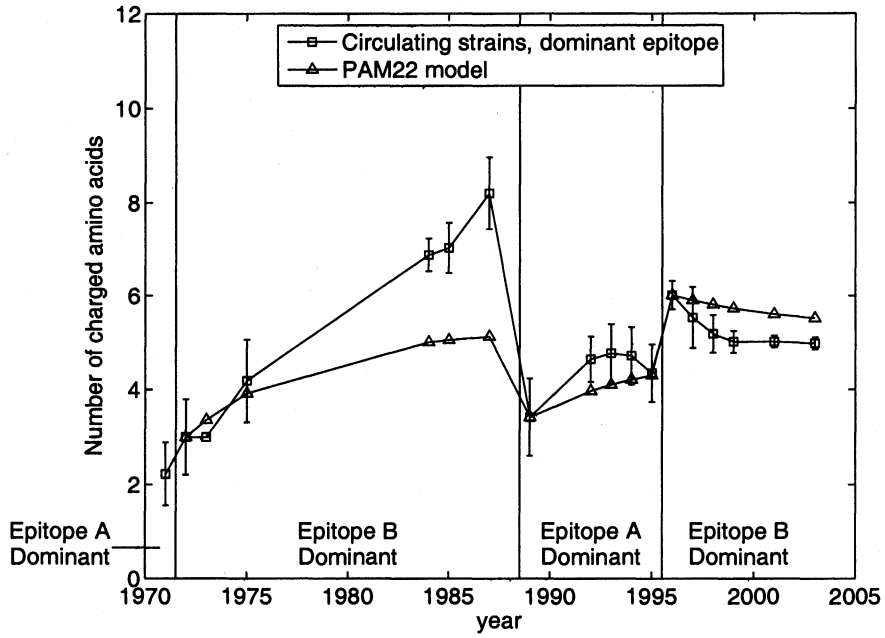


Figure 6.7 : Number of charged amino acids for each year on the dominant epitopes of the average of all the database strains in the same year. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 with PAM22 matrix were plotted. P_{c0} was fixed to the observed numbers of charged amino acids in the first year of the interval. The PAM22 matrix cannot reproduce the data. Error bars for the circulating strains are one standard deviation calculated from all NCBI strains collected in that year.

an inoculum of a dominate strain with the frequency 64.3%, as well as three closely-related variants with the frequencies 14.3%, 14.3%, and 7.1% respectively, totaling 35.7%, denoted CDC, or an inoculum of the dominant plaque-purified strain of the CDC stock mentioned above, denoted WyB4. The HA genes from replicated virus were sequenced from nasal washing samples at three days following intra-nasal inoculation. As observed in Table 6.2, naïve animals given the CDC inoculum resulted in an increase in the frequency of variants in the nasal wash virus. The WT:mutant ratio of 64.3%:35.7% in the CDC inoculum was reduced to 48.3%:51.7% in the progeny virus, indicating that the dominant viral strain in the starting material was reduced 16.0% while new HA mutants increased 16.0%. Moreover, most of the progeny variants characterized in the nasal wash samples were not previously detected in the CDC stock virus suggesting that the progeny variants arose during infection in the Guinea pig. The Guinea pig animal model also yielded new strains with multiple mutations as well as new strains which fundamentally changed the biochemical character of the mutated amino acid. Although many of the new variants may came from neutral mutation, the scenario that some of the new variants increased the replicative fitness cannot be precluded. Thus, it seems likely that the egg-adapted variants in the CDC stock had reduced replicative fitness relative to the WyB4 strain. Unexpectedly, reinfection of the same Guinea pigs with the CDC stock resulted in an increase of 11.3% in the frequency of the wildtype HA sequence and a concomitant equal decrease in the frequency of progeny strains having mutated HA genes.

In the second set of experiments in which naïve Guinea pigs were infected with the WyB4 plaque-purified clone of the dominant strain in the CDC stock, progeny virus had a 69.4%:30.6% WT:mutant HA gene ratio relatively close to that seen with the CDC stock in naïve animals. In the third experiments in which immunized animals

were infected with the WyB4 virus, the ratio of WT:mutant HA genes increased to 75.7%:24.3%, similar to the result observed in infections with the CDC stock. Taken together, infection of naïve animals resulted in a frequency of WT HA genes of approximately 65–70% while infection of immune animals produced an increase in the WT frequency to ~76% regardless of whether the CDC stock or WyB4 virus was used. By comparing the percentages of mutants listed in Table 6.2, three pairs of numbers are statistically significantly different ($p < 0.02$), which are CDC naïve vs. CDC reinfected, CDC naïve vs. WyB4 naïve, and CDC naïve vs. WyB4 HA-immunized, respectively. The null hypothesis cannot be rejected for the other three pairs when testing the statistically significant difference ($p > 0.3$). That is, only the low WT ratio in CDC naïve is statistically significant. Presumably the immunized animals eliminated the virus more quickly, and so fewer mutations were formed.

Figure 6.8 compares the number of charged residues predicted by both the PAM22 model and the reduced alphabet Markov model (RAMM), with those observed in the progeny of Guinea pigs infected with the A/Wyoming/2003 mixture obtained from the CDC. Because epitope B was the dominant epitope in the Wyoming virus circulating in 2003, we focused our analysis on the residues contained within this epitope. Five charged residues were present in the 21 amino acid B epitope in the initial Wyoming inoculum. The mean number of charged residues increased to 5.09 in the progeny from naïve animals and 5.09 in the progeny of reinfected animals.

The evolutionary times between the strains from the original inocula, the progeny of naïve Guinea pigs, and the progeny of reinfected animals are estimated by the mean number of mutations in the whole sequence of strains collected from Guinea pigs, divided by the approximate annual historical number of mutations in the whole HA1 sequences (mean historical mutations during the years 1971–2003 = 5.2 amino

acids/HA1 sequence/year). Note that regardless of whether charged amino acids are involved, the numbers of mutations in the whole sequence instead of epitope B are used to calibrate the evolutionary times. The strains from the naïve Guinea pig experiments were characterized, and the total number of point mutations in these strains was then divided by the number of strains to calculate the population average number of point mutations in the progeny strains. This average number was divided by 5.2, the mean number of historical mutations per year in the whole HA1 sequences, to obtain the evolutionary time Δt for the naïve Guinea pig experiment. The Δt for the reinfection experiment was calculated following the same method. The number of charged amino acids in epitope B, the dominant epitope for the Wyoming strain, increased in a trajectory that closely followed the path predicted by the RAMM model. In contrast the generic protein evolution PAM22 model predicted no such increase in charged residues. The evolutionary times and the numbers of charged amino acids in epitope B are calculated respectively by different approaches, therefore the evolutionary time is independent of the increase of charged amino acids. Thus, the agreement between RAMM model and the experiment is nontrivial.

Figure 6.9 shows a similar plot of charged residues observed from the experimental infection of naïve and immune Guinea pigs with the WyB4 isolate compared to predicted values. The methods used in the analysis shown in Figure 6.8 were applied to the data derived from infection with the homogenous virus. The WyB4 appeared to accumulate charged residues in the epitope B somewhat more rapidly than the Wyoming mixture obtained from the CDC. These data suggest either that the WyB4 strain has a replicative advantage in Guinea pigs and therefore evolves faster, or that accumulation of charged residues, especially in immune animals, represents a selective advantage.

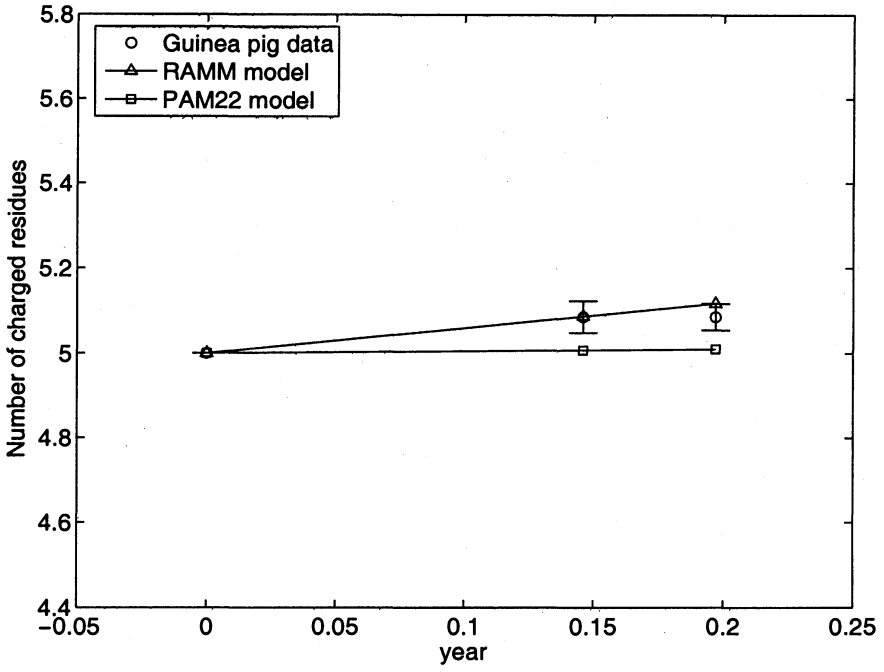


Figure 6.8 : Comparison of charged residue changes between theoretical models and sequence data derived from Guinea pigs inoculated with the CDC A/Wyoming/2003 virus mixture. The RAMM and the PAM theoretical models were considered, as well as three data points from the Guinea pig infections: First point: Wyoming inoculum; Second point: progeny strains from infection of naïve Guinea pigs; and Third point: progeny strains from infection of previously infected Guinea pigs. The time of naïve and reinfection strains was estimated from the average number of amino acid mutations, counting both wild type and mutated strains, in the whole HA1 sequence. With the assumption derived from historical data that the annual mutation rate was 5.2 amino acids/HA1 sequence/year, we divided those average numbers of mutations by 5.2 to obtain the times. Error bars are one standard error.

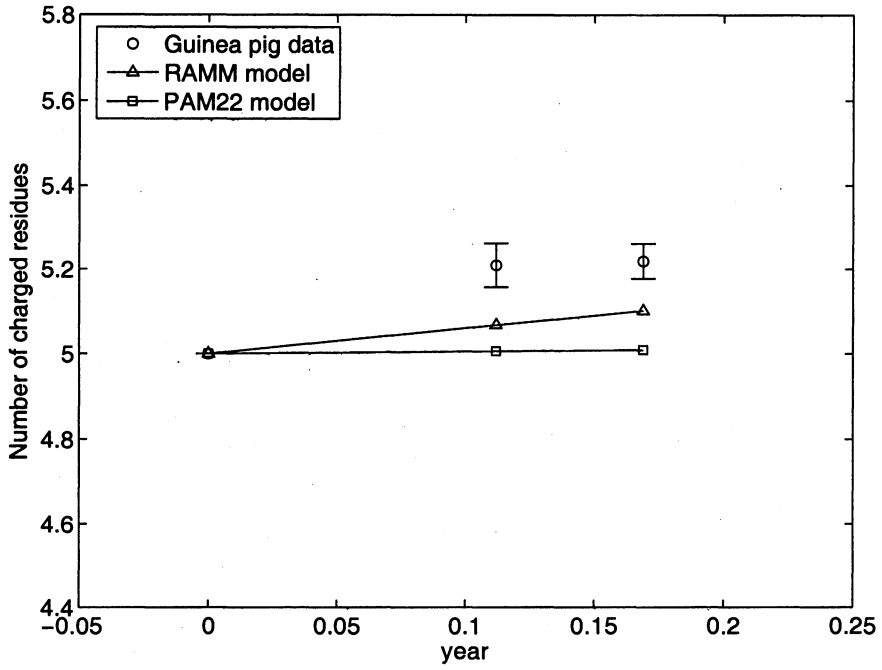


Figure 6.9 : Comparison of charged residue changes between theoretical models and sequence data derived from progeny virus isolated from Guinea pigs inoculated with the homogeneous WyB4 virus isolate. The number of predicted and observed charged residues were analyzed with the method used for the data in Figure 6.8. Error bars are one standard error.

Table 6.2 : Sequence analysis of progeny virus isolated from nasal washes of infected Guinea pigs. CDC virus designates mixture of Wyoming HA sequence variants contained in initial virus stock. WyB4 virus designates purified stock from predominant isolate of CDC virus. Immune status refers to whether the animals were naïve, i.e. neither previously infected nor immunized with purified HA protein, or immune. Number of sequences refers to the number of HA genes examined in progeny virus isolated from nasal washes.

Virus	Immune status	Number of sequences	Wild type	Mutants
CDC	naïve	58	28 (48.3%)	30 (51.7%)
CDC	reinfected	82	62 (75.6%)	20 (24.4%)
WyB4	naïve	62	43 (69.4%)	19 (30.6%)
WyB4	HA-immunized	210	159 (75.7%)	51 (24.3%)

Comparing Figure 6.8 and Figure 6.9, the WyB4 isolate appeared to evolve more in naïve animals than did the mixed CDC stock. A greater number of mutations per mutant strain were observed in naïve animals receiving the WyB4 isolate virus than the naïve animals receiving the CDC swarm. Among the 30 collected mutants of CDC naïve, the numbers of mutants with 1–4 point mutations were 22, 3, 4, and 1, respectively, with the average number of mutations per mutant equaled 1.47 (standard error: 0.16), and 15 of 44 (34.1%) mutations occurred in epitopes A–E. Among the 19 collected mutants of WyB4 naïve, the number of mutants with 1–4 point mutations were 6, 10, 2, and 1, respectively, with the average number of mutations per mutant equaled 1.89 (standard error: 0.19), and 21 of 36 (58.3%) mutations occurred in epitopes A–E. The difference of the average number of mutations per mutant of CDC naïve and WyB4 naïve is hence significant ($p = 0.039$). Again, the mechanisms behind these data are not clear, but they suggest the appearance of some form of

heterotypic immunity in the CDC stock virus. That is, the immune system of the Guinea pig either put more pressure on the WyB4 strain to evolve or took longer to clear the WyB4 and so allowed it to evolve longer relative to the CDC stock virus. An estimation of evolutionary time from the total number of mutations in the HA protein suggests the explanation is more likely the former.

6.3.4 Partitioning the Amino Acids by Charge is Optimal

The RAMM model discussed in this chapter deals with the numbers of amino acids belonging to a certain category in different years. Besides the charged–uncharged categorization of the 20 amino acids, we considered 6 different categories, which were polar (Arg, Asn, Asp, Cyc, Glu, Gln, His, Lys, Ser, Thr, Tyr), hydrophobic (Ala, Gly, Ile, Leu, Met, Phe, Pro, Trp, Val), basic (Arg, Lys, His), acidic (Asp, Glu), amides (Asn), aliphatic (Gly, Ala, Val, Leu, Ile), to generate different data sets for the RAMM model. Four combinations among these seven categories were also introduced with the amino acids in each category being counted once, which were charged+basic, charged+acidic, charged+polar, hydrophobic+aliphatic, respectively. To further generalize the R^2 value given by the RAMM model, different combinations of epitopes were focused on. Here the sequence of H3 hemagglutinin (HA1) between position 44 and 312 was split into six subsets: epitope A–E, and the amino acids outside any epitope, defined as O. Table 6.3 shows that charged amino acids are among the categories best fit by the RAMM model. Table 6.4 lists the R^2 values of the RAMM model for 7 categories of amino acids that are counted in 6 combinations of epitopes containing epitope B, the dominant epitope in the year of 1972–1987. Compared with Table 6.3, Table 6.4 indicates that focusing on the dominant epitope leads to the approximately optimal fitting.

Table 6.3 : R^2 values for amino acid categories and epitope B, the dominant epitope, using dominant circulating strains in 1972–1987 from which the model is trained. Seven categories and four combinations of categories are presented here. These four combinations involving charged and hydrophobic amino acids are chosen as the supplement to the seven categories, because charged and hydrophobic amino acids, especially charged ones, are critical in protein-protein interaction and evolution [8, 9, 10, 11, 12].

Categories	R^2	Categories	R^2
Charged	0.9810	Charged+Basic	0.9932
Polar	0.9620	Charged+Acidic	0.9567
Hydrophobic	0.9586	Charged+Polar	0.9568
Basic	0.9872	Hydrophobic+Aliphatic	0.8528
Acidic	0.8743		
Amides	0.6217		
Aliphatic	0.7698		

Table 6.4 : R^2 values for amino acid categories and combinations of epitopes involving epitope B, the dominant epitope, using dominant circulating strains in 1972–1987 from which the model is trained. Residues outside an epitope are denoted by O. The dominant circulating strains in the time span of 1972–1987 have epitope B as the dominant epitope. The H3N2 virus emerged in 1968, therefore less adaptation to host immune system was developed compared with other time span. The previous discuss indicates that epitope B had the immunodominance in this period of time, and this table shows the limited contribution of amino acids outside the dominant epitope to the pattern of evolution.

Epitopes	AB	BC	BD	BE	BO	ABCDEO
Charged	0.8928	0.9782	0.9431	0.9913	0.9810	0.8935
Polar	0.8187	0.9620	0.7780	0.8528	0.9766	0.7656
Hydrophobic	0.9586	0.9771	0.7928	0.9580	0.9884	0.9766
Basic	0.9872	0.9872	0.9913	0.9456	0.9872	0.9917
Acidic	0.4730	0.9210	0.6337	0.7245	0.6838	0.3916
Amides	0.9771	0.7922	0.9834	0.2171	0.8443	0.9326
Aliphatic	0.4976	0.9009	0.7648	0.3787	0.7895	0.6129

6.4 Discussion

6.4.1 Data Fitting and Verifying the RAMM Model

For the historical data, our analysis was based on both the dominant circulating strain and the average of all circulating strains deposited in NCBI database. The frequency of database strains is not currently available. In addition, because clear sequential and temporal relationships between the sequences are unavailable, the fixation of mutations is unknown. Although some trends in the mutations have emerged (such as increased numbers of both charged residues in defined epitopes and total numbers of glycosylation sites throughout the HA ectodomain), the majority of the sequences do not appear to be fixed.

For the Guinea pig animal model, the frequency of both wildtype and variants were collected and tabulated. These data allowed us to calculate a population average of all the strains with no loss of frequency information when counting the average number of mutations and the increase of charged amino acids. For the Guinea pig experiment, we have sequences for all the strains characterized in all animals, and can thus calculate the true population average. However, at this time we do not have firm data on fixation rates of variants in the Guinea pig infection. Some of the observed variants could have been fixed over the 4–6 rounds of replication that occurred in the three-day infection, but a serial transmission experiment performed in immune animals would be required to provide quantitative data on mutations have become immunologically fixed. Thus, the quality of the Guinea pig data and the human database with respect to fixation of mutations is comparably similar.

Note the RAMM model calibrated to years of human epidemiological data fit well the Guinea pig experimental data. Figure 6.8 and Figure 6.9 compare the observed

evolution of the HA envelope in the CDC Wyoming mixed stock and the purified WyB4 isolate with the evolution predicted by the two models. As shown above, the model fit the in vivo data well for the CDC stock but underestimated the accumulation of charged residues observed in progeny from the WyB4 infections. CDC stock was an ensemble of a dominant and three closely-related variants while WyB4, the plaque-purified isolate and the dominant member of the CDC stock was the virus that appeared to have greater replicative capacity in Guinea pigs as compared to the variants in the CDC stock. The CDC stock infection of naïve animals more closely represents a cross-section of a transmitting quasi-species related to all the circulating strains existing in one year, while the WyB4 infection of naïve animals may model the dominant circulating strain.

Table 6.4 provides further support to the hypothesis of a single dominant epitope. These data show the R^2 values of the correlations between observations and the model using numbers of charged amino acids in a variety of combinations of epitopes as well as a general null model using the whole sequence of H3 HA. Different amino acid categories are also incorporated in Table 6.4 to show the R^2 values of the correlation between observations and the model. Thus, as in previous studies [61], Table 6.4 confirms that calculations using the single dominant epitope lead to the approximately optimal fitting between the observations and the model. None of the inclusions of subdominant epitopes yields R^2 values much greater than that of focusing on the single dominant epitope.

The results obtained from the discrete–time and continuous–time Markov chains were essentially identical. By comparing expressions (6.3) and (6.7), the parameter set of the discrete–time Markov chain $(\epsilon, \delta, P_{c0})$ and that of the continuous–time Markov

chain $(\lambda_1, \lambda_2, P_{c0})$ are related by

$$\begin{cases} \epsilon &= 1 - e^{-(\lambda_1+\lambda_2)} \\ \delta &= \dfrac{\lambda_2}{\lambda_1+\lambda_2} - \dfrac{1}{4} \end{cases} . \tag{6.17}$$

The discrete–time Markov chain was appropriate for the case in this chapter because the time unit for the data was year. Currently there is limited time resolution available, hence using discrete–time model with one-year time resolution did not lose any information. The virus evolves continuously in different geographic regions, however, and spreads among these regions by the migration of humans and birds. Furthermore, the global virus strains in a given time point constitutes an ensemble where the population of different strains varies. Available data are the dominant circulating strains in each year, as well as the database strains. Therefore, the average numbers of charged amino acids should be the weighted sum of all the database strains in each year, rather than their simple arithmetic means. It will be of significant help to know the frequency of each database strain so that a population average can be calculated.

6.4.2 Model Reversibility

Most previous models assumed the reversibility of the Markov chain, whose transition matrices $\mathbf{P} = \{P_{ij}\}$ and equilibrium probability distributions $\boldsymbol{\pi} = \{\pi_i\}$ satisfied the detailed balance equation

$$\pi_i P_{ij} = \pi_j P_{ji}. \tag{6.18}$$

Such models were mainly designed to describe the evolution of bacteria, archaea, and eukaryota living in a relative constant environment. The evolution of these species is probably dominated by a stable and loose natural selection that is tolerant to mildly deleterious polymorphism [151]. Viruses, on the other hand, perform their biological function in other organisms, and usually their surface protein antigens are

targeted by the immune systems in the antigen sequence space that makes the virus evolution under strong, directed, and changing selection. This arrow of time due to directed selection invalidates the reversibility assumption, because $\boldsymbol{\pi}$ is changing, and this higher degree of freedom must be accounted for when formulating the model. In the data shown here, the detailed balance condition (6.18) was not established, because the immune pressure and selection are constantly changing. In other words, when starting from a non-steady state initial condition, the system will first converge toward the steady state before fluctuating about that steady state until the immune pressure changes again. We studied the non-equilibrium dynamics of the evolution: the evolution of the probability distribution was dominated by a changing proportion of charged amino acids.

Compared with Figure 6.4 and Figure 6.5, PAM22 fit better with both the circulating strains and the database in later years than in early years partly because the equilibrium probability distribution of PAM22 model (6.16) solved by (6.18) was $P_c = 0.25, P_u = 0.75$, hence the expected numbers of charged amino acid in epitope A and epitope B were 5.31 and 4.81, close to the numbers in later years depicted in Figure 6.1 and Figure 6.3. On the other hand, PAM22 drifted away from both the circulating strains and the database strains in the early stage after the emergence of H3N2 virus, when the immune escape was underway most strongly. This phenomena showed that H3N2 virus did not follow the general law of protein evolution in a constant environment in the early years, while its evolution returned to steady state about twenty years after its emergence in 1968. The higher fixation rate of charged residues determined by the RAMM model versus the general PAM model is due to immune pressure.

6.4.3 Fluctuation and Spatial Distribution of Charge

The dominant epitopes [2] in history were epitope A and epitope B, so in a certain year one of them was dominant while the other was subdominant. As a common trend, the number of charged amino acids on the dominant epitope increased and that on the subdominant epitope decreased. The explanation is likely that an increase in number of charged amino acids reduces the free energy of binding of antibodies to the epitope, by increasing the affinity of the epitope for water. Although the number of charged amino acid on the subdominant epitope decreased almost monotonically, it never decreased to the initial level.

[9] pointed out that the epitopes of A/Aichi/2/1968 and A/Kamata/14/1991 share similar shapes, hence the spatial distribution of charged amino acids on the epitopes could be estimated from the available structure of Aichi strains in 1968 (PDB entry: 1KEN). We plotted such figures for epitope A and B for each circulating strain. The number and position of hydrophobic amino acids were relatively conservative: hydrophobic amino acids existed at position (130, 138, 168) on epitope A and position (163, 194, 196, 198) on epitope B in each circulating strain. Positions (130, 138, 168) located in three corners of epitope A while positions (163, 194, 196, 198) constituted a continuous region on the center of epitope B. The hydrophobic center on epitope B was gradually surrounded by charged amino acids during the virus evolution in 1972–1987, so in the epitope shift year of 1989, epitope B contained a hydrophobic center surrounded by charged amino acid, a structure found in other protein-protein interaction cases [8]. This arrangement of amino acids was preserved in epitope B in later years, which may be the reason that the number of charged amino acids stayed in a high level in epitope B. Epitope A, however, did not evolve into the stable structure with a hydrophobic center surrounded by charged residues; thus the number

of charged amino acids fell to a lower level when epitope A was subdominant than when epitope B was subdominant. We also repeated the same analysis on the recently emerged H5N1 strains, using as epitopes the residues aligned to H3N2 epitopes. We did not observe a significant change in the number of charged amino acids, perhaps since the H5N1 virus has not evolved substantially in human.

6.5 Conclusion

Mutation of the H3 hemagglutinin on the surface of the the influenza virus decreases the ability of the immune system to recognize the flu and decreases the efficacy of the annual vaccine. We show that influenza tends to increase the number of charged amino acids in the regions of hemagglutinin that the immune system recognizes, probably because this reduces ability of antibodies to bind hemagglutinin. An interesting corollary of this selection is that the number of charges in the dominant epitope of the dominant circulating virus strain is never fewer than that in the vaccine strain, chosen early in the season. We developed a model of the evolution of charge in hemagglutinin by partitioning 20 amino acids into two categories: charged and uncharged, calibrated this model on virus evolution data in humans, and demonstrated the model on Guinea pig animal model studies.

Protein evolution models such as PAM model and BLOSUM model typically apply to the evolution of bacteria, archaea, and eukaryota. For influenza virus, the harsh and changing environment due to immune pressure on the virus, makes its evolution a non-equilibrium dynamics, especially in the time period after its initial emergence in humans. The RAMM model supports the hypothesis that the rate of charge evolution is greater in regions of hemagglutinin recognized by the immune system than in proteins in general. Such temporal and spatial heterogeneity requires a method such

as we have presented here for modeling the virus evolution.

6.6 Supplementary Material

6.6.1 Fitting the Markov Models

Two methods were applied to fix the parameters: a non-linear least square fitting and the maximum likelihood estimation. To fix the data points for the least square fitting, the observed numbers of charged amino acids N_c were asserted to be equal to the means of the binomial distribution each year. The parameters $\boldsymbol{\theta} = (\epsilon, \delta, P_{c0})$ in the discrete–time Markov chain were fit by non-linear least squares in each time interval based on the observed values $P_c(t \mid \epsilon, \delta, P_{c0}) = N_c/N$, using the parametric form (6.3).

The estimated number of annually mutated residues on epitope B in this interval was

$$\begin{aligned} \text{number of mutated residues} &= N_c\epsilon\left(1-\frac{1}{5}\alpha_{cc}\right)+N_u\epsilon\left(1-\frac{1}{15}\alpha_{uu}\right) \\ &= \frac{19}{20}N\epsilon+\epsilon\delta\left(\frac{N_u}{15}-\frac{N_c}{5}\right) \\ &\approx \frac{19}{20}N\epsilon \end{aligned} \tag{6.19}$$

where N_c and N_u were the numbers of charged and uncharged amino acid residues in the dominant epitope, respectively.

When ϵt is small, the original model of $P_c(t \mid \epsilon, \delta)$ reduced to

$$\begin{aligned} P_c(t \mid \epsilon, \delta) &= \frac{1}{4}\left(1+4\delta-(1-\epsilon)^t\left(1+4\delta-4P_{c0}\right)\right) \\ &\approx \frac{1}{4}\left(1+4\delta-(1-\epsilon t)\left(1+4\delta-4P_{c0}\right)\right) \\ &= P_{c0}+\epsilon\delta t \end{aligned} \tag{6.20}$$

which meant that when $\epsilon \to 0$ and $\epsilon\delta \to$ const, $\epsilon\delta$ was the rate of increase of the probability that a charged amino acid exists at a certain position on the dominant epitope.

Chapter 7

Predicting Fixation Tendencies of the H3N2 Influenza Virus by Free Energy Calculation

Influenza virus evolves to escape from immune system antibodies that bind to it. We used free energy calculations with Einstein crystals as reference states to calculate the difference of antibody binding free energy ($\Delta\Delta G$) induced by amino acid substitution at each position in epitope B of the H3N2 influenza hemagglutinin, the key target for antibody. A substitution with positive $\Delta\Delta G$ value decreases the antibody binding constant. On average an uncharged to charged amino acid substitution generates the highest $\Delta\Delta G$ values. Also on average, substitutions between small amino acids generate $\Delta\Delta G$ values near to zero. The 21 sites in epitope B have varying expected free energy differences for a random substitution. Historical amino acid substitutions in epitope B for the A/Aichi/2/1968 strain of influenza A show that most fixed and temporarily circulating substitutions generate positive $\Delta\Delta G$ values. We propose that the observed pattern of H3N2 virus evolution is affected by the free energy landscape, the mapping from the free energy landscape to virus fitness landscape, and random genetic drift of the virus. Monte Carlo simulations of virus evolution are presented to support this view.

7.1 Introduction

Influenza A virus causes annual global epidemics resulting in 5–15% of the population being infected, 3–5 million severe cases, and 250,000–500,000 fatalities [14]. The

subtype of influenza A is determined by two surface glycoproteins—hemagglutinin (H) and neuraminidase (N). The H3N2 virus has been one of the dominant circulating subtypes since its emergence in 1968. The antibodies IgG and IgA are the major components of the immune system that control influenza infection, binding to the influenza hemagglutinin [72]. There are five epitopes at the antibody binding sites on the top of H3 hemagglutinin, namely epitopes A–E. The epitope bound most prolifically by antibody is defined as the dominant epitope, and it is central to the process of virus neutralization by antibody and virus escape substitution [2]. The cellular immune system, on the other hand, plays a relatively less recognized role in handling the invasive influenza virus [72]. The cellular system along with the innate immune system exerts a somewhat more homogeneous immune reaction against genetically distinct influenza strains [73, 72].

Vaccination is currently the primary method to prevent and control an influenza epidemic in the human population [14]. Influenza vaccination raises the level of antibody specific for hemagglutinin and significantly enhances the binding affinity between antibody and hemagglutinin. Vaccine effectiveness depends on the antigenic distance between the hemagglutinin of the administered vaccine strain and that of the dominant circulating strain in the same season [2, 74]. Memory immune response from virus in previous seasons as well as vaccination in the current and previous seasons impose selective pressure on the current circulating virus to force it to evolve away from the virus strains recognized by memory antibodies that selectively bind to hemagglutinin.

As a result of the immune pressure and the escape evolution of the influenza virus, which is largely substitution in the dominant epitope of hemagglutinin, the influenza vaccine must be redesigned and administered each year, and the vaccine effectiveness

has been suboptimal in some flu seasons [2, 33]. The escape evolution in the dominant epitope is at a higher rate than that in the amino acid sites outside the dominant epitope [42]. Sites in the dominant epitope also show higher Shannon entropy of the 20 amino acids than do those outside the dominant epitope [1]. High substitution rate and Shannon entropy in the dominant epitope of hemagglutinin suggest that the dominant epitope is under the strongest positive selection by human antibodies. The immune pressure against each genotype of the dominant epitope can be at least partially quantified by the binding constant between antibody and hemagglutinin.

The H3N2 virus and human immune system in this work are simplified to be a system consisting of the H3 hemagglutinin and the corresponding human antibody. Exposure by infection or vaccination produces an affinity-matured antibody with the binding constant to the corresponding hemagglutinin equal to 10^6–10^7 M^{-1}, while the binding constant of an antibody uncorrelated to the hemagglutinin is below 10^2 M^{-1} [72]. Escape substitutions may decrease the binding constant by changing the antibody binding free energy ΔG. Some substitutions decrease the antibody binding constant more than others and have higher probabilities to be fixed, because decrease in the antibody binding constant is favorable to the virus. Here we define the difference of antibody binding free energy as $\Delta\Delta G = \Delta G_{42} - \Delta G_{31}$ in which ΔG_{31} and ΔG_{42} are antibody-wildtype hemagglutinin binding free energy and antibody-evolved hemagglutinin binding free energy, respectively, as shown in Figure 7.1. The fixation tendency of each substitution is a function of the difference of the antibody binding free energy [75] of the escape substitution.

Epitope A or B of the H3N2 virus was dominant in most influenza seasons [2]. Epitope B of the H3N2 virus was the dominant epitope presenting more substitutions than any other epitope in the recent years. Epitope B was also dominant in 1968

when H3N2 virus emerged. Thus during these periods of time, the substitutions in epitope B directly affect the antibody binding constant and reflect the direction of the virus escape substitution. To attain a global view of the effects of substitutions in epitope B, it is necessary to compute a matrix containing the differences of antibody binding free energy caused by each possible single substitution in epitope B. There are 21 amino acid sites in epitope B, and each residue in the wild type strain may substitute to any of the 19 different types to amino acid residues, hence we need to calculate a 19×21 matrix with 399 elements. Such a matrix is a free energy landscape quantifying the immune selection over each evolved influenza strain. In this free energy landscape, the virus tends to evolve to a position with low binding affinity of antibody to evade antibodies and reduce the immune pressure. Calculation of this landscape will enable us to study the mechanism of immune escape from a quantitative viewpoint, providing a criterion to describe and foresee the evolution of influenza virus.

This chapter is organized as follows: In Materials and Methods section, we describe the protocol for the free energy calculation and the system of hemagglutinin and antibody. In Results section, we present and analyze the calculated free energy landscape. The substitutions observed in history are also compared with the results of the calculation. In the Discussion section, a general picture of H3N2 virus evolution under the selection pressure of the immune system is discussed and simulation results are discussed. Finally, our work is summarized in the Conclusion section.

7.2 Materials and Methods

7.2.1 Scheme of the Free Energy Calculation

The expression of the binding constant K depends on the antibody binding free energy ΔG, $K = \exp(-\Delta G/RT)$. The Boltzmann constant $R = 1.987 \times 10^{-3}$ kcal/mol/K. The temperature is fixed to the normal human body temperature $T = 310$ K. Shown in Figure 7.1, one substitution in hemagglutinin changes the antibody binding free energy from ΔG_{31} to ΔG_{42}. The first and second subscripts define the end state and the starting state of the binding process, respectively. The ratio of the antibody binding constant after and before substitution is written as

$$\frac{K_1}{K_0} = \exp(-\Delta\Delta G/RT) \tag{7.1}$$

where K_1 and K_0 are the antibody binding constant to substituted and wildtype hemagglutinin, respectively.

The difference of the antibody binding free energy $\Delta\Delta G = \Delta G_{42} - \Delta G_{31} = \Delta G_{43} - \Delta G_{21}$ is calculated by applying the Hess' Law to the thermodynamic cycle defined by State 1–4 in Figure 7.1. The processes corresponding to ΔG_{43} and ΔG_{21} are unphysical but more convenient to simulate. We calculated ΔG_{21} and ΔG_{43} for each amino acid substitution in the unbound hemagglutinin and hemagglutinin bound by antibody, respectively. On the surface of the virus particle, hemagglutinin exists in the form of a trimer in which three monomers are encoded by the same virus gene. Thus we simultaneously substituted the amino acids in three hemagglutinin monomers in the trimer. The antibody has a Y-shaped structure with two heavy chains and two light chains. In the resolved structure (PDB code: 1KEN), the hemagglutinin trimer is bound by two Fab fragments. Thus, we incorporated the Fab dimer into the system for MD simulation.

Using the software CHARMM [152], we calculated each of ΔG_{21} and ΔG_{43} using thermodynamic integration [153]. We used molecular dynamics (MD) simulation to obtain the ensemble averages of the integrand from which each of ΔG_{21} and ΔG_{43} is calculated. The potential energy for the MD algorithm to sample the conformation space of the system is

$$U(\boldsymbol{r}, \lambda) = (1 - \lambda) U_{\text{reac}}(\boldsymbol{r}) + \lambda U_{\text{prod}}(\boldsymbol{r}) \tag{7.2}$$

in which $\boldsymbol{r}$ is the coordinates of all the atoms, λ is the variable of integration, U_{reac} is the potential energy of the system corresponding to wildtype hemagglutinin, and U_{prod} is the potential energy of the system corresponding to substituted hemagglutinin. The value of ΔG_{21} or ΔG_{43} is

$$\Delta G = \int_0^1 \left\langle \frac{\partial U(\boldsymbol{r}, \lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda = \int_0^1 \langle U_{\text{prod}}(\boldsymbol{r}) - U_{\text{reac}}(\boldsymbol{r}) \rangle_\lambda \, \mathrm{d}\lambda. \tag{7.3}$$

The integrand $\langle U_{\text{prod}}(\boldsymbol{r}) - U_{\text{reac}}(\boldsymbol{r}) \rangle_\lambda$ is the ensemble average with fixed λ of potential energy difference between the system after and before substitution. The interval of integration $\lambda \in (0, 1)$ was equally divided into four subintervals in each of which a 16-point Gauss-Legendre quadrature was applied to numerically integrate the ensemble averages. The ensemble averages with 64 distinct $\lambda \in (0, 1)$ were calculated by MD simulation with the potential energy defined in equation 7.2.

7.2.2 Einstein Crystal

We introduce the Einstein crystals to calculate the free energy of the reference state in the dual topology at both endpoints of the thermodynamic integration. To illustrate the function of the Einstein crystals, we analyze the free energy of the dual topology without Einstein crystals when $\lambda = 0$ as an example. We denoted by n_1, n_2, and n_0 numbers of the reactant atoms, product atoms, and all the remaining atoms in the

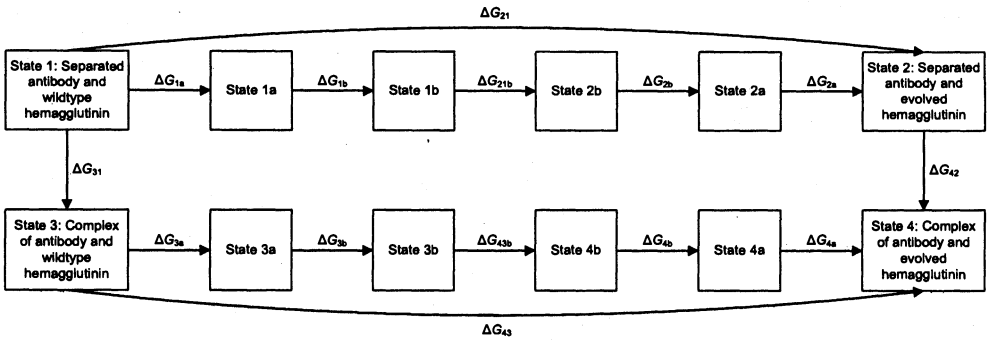


Figure 7.1 : The scheme of the free energy calculation. The free energy difference of one substitution is calculated by $\Delta\Delta G = \Delta G_{43} - \Delta G_{21}$. State n, $n = 1$–4, is the real system. State na has the same configuration of atoms as state n except that all the hydrogen atoms have mass 16.000 amu. Compared to state na, state nb contains one additional Einstein crystal of product atoms ($n = 1, 3$) or reactant atoms ($n = 2, 4$). The mass of hydrogen atoms in state nb is also 16.000 amu. Free energy $\Delta G_{21\mathrm{b}}$ and $\Delta G_{43\mathrm{b}}$ are obtained by thermodynamic integration.

system, respectively. We denoted by $\boldsymbol{r}$, $\boldsymbol{r}_{\mathrm{product}}$, and $\boldsymbol{x}$ the coordinates of the reactant atoms, product atoms, and all the remaining atoms in the system. The momenta of reactant atoms, product atoms, and the remaining atoms are denoted by $p_{\mathrm{r},i}$, $p_{\mathrm{p},i}$, and $p_{\mathrm{x},i}$. The masses are similarly denoted by $m_{\mathrm{r},i}$, $m_{\mathrm{p},i}$, and $m_{\mathrm{x},i}$. The Hamiltonian of the system with $\lambda = 0$ is

$$H = \sum_{i=1}^{n_0} \frac{p_{\mathrm{x},i}^2}{2m_{\mathrm{x},i}} + \sum_{i=1}^{n_1} \frac{p_{\mathrm{r},i}^2}{2m_{\mathrm{r},i}} + \sum_{i=1}^{n_2} \frac{p_{\mathrm{p},i}^2}{2m_{\mathrm{p},i}} + U_{n_0}(\boldsymbol{x}) + (1-\lambda)\, U_{n_0+n_1}(\boldsymbol{x}, \boldsymbol{r}) + \lambda U_{n_0+n_2}(\boldsymbol{x}, \boldsymbol{r}_{\mathrm{product}}). \tag{7.4}$$

The partition function is

$$\begin{aligned} Q =& \prod_{i=1}^{n_0} \left(\frac{2\pi m_{\mathrm{x},i}}{h^2\beta}\right)^{3/2} \prod_{i=1}^{n_1} \left(\frac{2\pi m_{\mathrm{r},i}}{h^2\beta}\right)^{3/2} \prod_{i=1}^{n_2} \left(\frac{2\pi m_{\mathrm{p},i}}{h^2\beta}\right)^{3/2} \\ & \int \mathrm{d}\boldsymbol{x}\ \mathrm{d}\boldsymbol{r}\ \exp\left[-\beta U_{n_0}(\boldsymbol{x}) - \beta U_{n_0+n_1}(\boldsymbol{x},\boldsymbol{r})\right] \times \int \mathrm{d}\boldsymbol{r}_{\mathrm{product}}\ 1 \\ =& Q_{\mathrm{real}} \times \prod_{i=1}^{n_2} \left[\left(\frac{2\pi m_{\mathrm{p},i}}{h^2\beta}\right)^{3/2} V\right] \\ =& Q_{\mathrm{real}} \times Q_{\mathrm{product}}. \end{aligned} \tag{7.5}$$

When $\lambda = 0$, this partition function is the product of Q_{real}, the partition function of the real system without product atoms, and Q_{product}, the partition function of the product atoms when $\lambda = 0$.

The free energy is given by $-1/\beta$ times the logarithm of the above partition function. The free energy is

$$G = G_{\mathrm{real}} - \frac{1}{\beta} \sum_{i=1}^{n_2} \frac{3}{2} \log\left(\frac{2\pi m_{\mathrm{p},i}}{h^2\beta}\right) - \frac{1}{\beta} n_2 \log V. \tag{7.6}$$

As shown in the above equation, the effect on the translational entropy from the product atoms is proportional to the logarithm of system size V. It diverges in the thermodynamic limit. This divergence exists, no matter what λ scaling is performed. Note that we do not use the Einstein crystals to handle the translational entropy a ligand loses or gains when binding a flexible biomolecular receptor, which is taken into account by the thermodynamic cycle in Figure 7.1. The translational entropy, proportional to $\log V$ in equation 7.6, is that of the dummy product atoms, not that of the bound or unbound complex.

The value of G depends on the identity of the product atoms. Thus, the contribution to the thermodynamic integration is different at the two endpoints, i.e. $-kT \log Q_{\mathrm{reactant}} \neq -kT \log Q_{\mathrm{product}}$, in which Q_{reactant} is the partition function of the

reactant atoms when $\lambda = 1$. Note also that the expression of the partition function contains the factor Q_{product} for the product atoms. Relating the conventional expression for thermodynamic integration, equation 7.3, to $\Delta\Delta G$ of equation 7.1 requires one to account for this term. This term arises from the use of a dual topology in CHARMM, and this term is typically ignored. While the contribution from the decoupled atoms is not constant, it can be exactly calculated if the restricted partition function over the decoupled atoms can be calculated. This calculation is what the Einstein crystal performs, using an Einstein crystal for the reference state rather the ideal gas in equation 7.4.

In four 16-window thermodynamic integrations, the smallest variable of integration is $\lambda = 1.32 \times 10^{-3}$. Since λ is close to zero, product atoms in the system have potential energy near zero and behave as ideal gas atoms, with translational entropy proportional to the logarithm of system size, see equation 7.6. Exact calculation of the translational entropy terms of product atoms at $\lambda = 0$ by explicit dynamics seems difficult, because the translational entropy of the product atoms grows as the logarithm of the system size. These relatively free product atoms destabilize the system. This entropy divergence is a fundamental feature of the statistical mechanics, not a numerical artifact. Unrestrained product atoms induce large fluctuation of the Hamiltonian in the MD algorithm. These fluctuations increase the standard error of the quantity $U_{\text{prod}}(r) - U_{\text{reac}}(r)$, which is defined in equation 7.3 and is computed from the trajectory of the MD simulation. These fluctuations often cause the numerical integration algorithm in the MD simulation to be unstable [154]. In this case, the energy of the simulated system increases rapidly. This phenomenon causes CHARMM to terminate abnormally. The translational entropy introduced by the free atoms at $\lambda = 0$ and 1 affects the result. Reactant atoms cause the same problem near $\lambda = 1$.

We noticed that the non-linear scaling, i.e. using a high power of λ such as the fourth power of λ, in equation 7.2 [155, 156] did not work. The high power of the smallest λ is extremely close to zero and the product atoms are almost free, which cause the MD simulation to terminate abnormally at several windows with small λ. Additionally, the issue of translational entropy of reactant and product atoms needs to be addressed. Even when the MD algorithm with the non-linear scaling of λ [155, 156] terminates and appears to have generated a converged simulation trajectory, this does not necessarily imply that the translational entropy of reactant or product atoms has been properly controlled. In fact, the λ scaling approach may hide the entropy divergence at $\lambda = 0$ or $\lambda = 1$ by letting the algorithm terminate due to numerical roundoff error, rather than building statistical mechanical reference states for each of $\lambda = 0$ and $\lambda = 1$ to account for or control the effect of translational entropy.

An alternative to λ scaling introduces the soft-core potential as a way to turn off the potential [157, 158]. The soft-core approach, like the lambda-scaling approach, does not address the translation entropy of the atoms at $\lambda = 0$ or $\lambda = 1$. Previous studies with non-constrained atoms at both endpoints have been performed [159, 160, 161, 162, 163, 164, 165]. Besides the classical molecular dynamics with a non-ideal-gas reference state introduced into the dual topology, quantum molecular dynamics via metadynamics has been used to analyze a deamidation process [166]. Other applications of quantum molecular dynamics based free energy calculation include chorismate conversion to prephenate [167], isomerization of glycine [168], and histone lysine methylation [169]. As illustrated in equation 7.6, the translational entropy of the uncoupled atoms causes error in the final free energy results if it is not accounted for.

One way to calculate the free energy change exactly is to use a non-ideal-gas reference state. This is quite natural, since the protein is not composed of ideal gas atoms. Deng and Roux introduced restraint potentials to confine the translational and rotational motion of a bound ligand to accelerate convergence of the simulation [170]. We use this idea to exactly include the contribution from the restrained states and built two Einstein crystals as the reference states for reactant and product atoms, respectively. Our calculation allows a theoretically exact determination of the free energy due to amino acid substitution.

To handle these two difficulties at both endpoints of the integration in a theoretically exact way, we use two Einstein crystals as the reference states for reactant and product atoms, respectively. The Einstein crystal has been used as a reference state for free energy calculations. Frenkel and Ladd computed free energy of solids by building a path connecting the real solid and the reference Einstein crystal [171]. Noya et al. showed that a restrained Einstein crystal is a suitable reference in the free energy calculation of biomolecules [172]. The Einstein crystal, a solid state model, is consistent with the nature of antibody binding process in liquid phase. First, although the importance of biomolecular flexibility in protein-protein binding process is well-accepted, and is fully and exactly included in our calculation, we simply need to localize the product atoms when $\lambda = 0$ and the reactant atoms when $\lambda = 1$. Moreover, we need to calculate the contribution to the free energy of these localized atoms.

The choice of Einstein crystals as the reference states removes the singularity in thermodynamic integration in equation 7.3. As an example, an Einstein crystal was used as the reference state for the free energy calculation of hard-sphere fluid in order to remove the singularity in equation 7.3 at the end point $\lambda = 0$ [173]. In this example,

the reference Einstein crystal was achieved by harmonically coupling the particles to their equilibrium positions and removing all interactions between particles [174].

We here use Einstein crystals as the reference states to calculate the binding free energy change due to amino acid substitution. The Einstein crystal is a model for localized atoms. The free energy of the Einstein crystal can be exactly calculated. One Einstein crystal contains distinguishable and non-interacting atoms under harmonic constraints around reference positions fixed in space. In the Einstein crystal, the atom i with coordinates $\boldsymbol{r}_i$ has potential energy

$$U_i(\boldsymbol{r}_i) = \frac{K_i}{2} \|\boldsymbol{r}_i - \boldsymbol{r}_{i0}\|^2 \tag{7.7}$$

in which $\boldsymbol{r}_i$ and $\boldsymbol{r}_{i0}$ are the actual and reference position of the atom, respectively, and K_i is the force constant of the harmonic constraint. We denote by m_i the mass of atom i. The canonical partition function of an Einstein crystal is

$$\begin{aligned} Q_{\mathrm{E}}(N,V,T) &= \frac{1}{h^{3N}} \int \exp\left(\sum_{i=1}^{N} \frac{-\beta \boldsymbol{p}_i^2}{2m_i}\right) \exp\left(\sum_{i=1}^{N} \frac{-\beta K_i \|\boldsymbol{r}_i - \boldsymbol{r}_{i0}\|^2}{2}\right) \mathrm{d}\boldsymbol{p}\, \mathrm{d}\boldsymbol{r} \\ &= \left(\frac{2\pi}{h\beta}\right)^{3N} \prod_{i=1}^{N} \left(\frac{m_i}{K_i}\right)^{3/2}. \end{aligned} \tag{7.8}$$

The spatial fluctuation of atom i in the Einstein crystal is

$$\left\langle (\delta \boldsymbol{r}_i)^2 \right\rangle = \frac{3}{\beta K_i}. \tag{7.9}$$

In our system, we let the potential energy for MD simulation defined by equation 7.2 become

$$U(\boldsymbol{r},\lambda) = (1-\lambda)\, U_{\mathrm{reac}}(\boldsymbol{r}) + \lambda U_{\mathrm{prod}}(\boldsymbol{r}) + \lambda U_{\mathrm{ein,reac}}(\boldsymbol{r}) + (1-\lambda)\, U_{\mathrm{ein,prod}}(\boldsymbol{r}). \tag{7.10}$$

Therefore reactant and product atoms are localized at both $\lambda = 0$ and $\lambda = 1$. The reference positions of atoms in Einstein crystals are the equilibrium positions of corresponding reactant and product atoms. To minimize the numerical error during

the thermodynamic integration calculation, we minimized the fluctuation of the integrand of thermodynamic integration $\langle \partial U(\boldsymbol{r},\lambda)/\partial\lambda \rangle_\lambda = \langle U_{\text{ein,reac}}(\boldsymbol{r}) - U_{\text{reac}}(\boldsymbol{r}) \rangle_\lambda + \langle U_{\text{prod}}(\boldsymbol{r}) - U_{\text{ein,prod}}(\boldsymbol{r}) \rangle_\lambda$. Minimization of the terms on the right hand size is approximately achieved by letting the average spatial fluctuation of each atom in Einstein crystals equal to that of the corresponding reactant or product atom, i.e.

$$\left\langle (\delta \boldsymbol{r}_i)^2 \right\rangle_{\text{reac}} = \left\langle (\delta \boldsymbol{r}_i)^2 \right\rangle_{\text{ein,reac}} = \frac{3}{\beta K_i^{\text{reac}}} \tag{7.11}$$

$$\left\langle (\delta \boldsymbol{r}_i)^2 \right\rangle_{\text{prod}} = \left\langle (\delta \boldsymbol{r}_i)^2 \right\rangle_{\text{ein,prod}} = \frac{3}{\beta K_i^{\text{prod}}} \tag{7.12}$$

For each atom in the Einstein crystal, the force constant of harmonic constraint, K_i^{reac} or K_i^{prod}, was calculated from the monitored fluctuations of the corresponding reactant or product atom with equation 7.11 or equation 7.12. In the scheme in Figure 7.1, the states with Einstein crystals are states 1b, 2b, 3b, and 4b.

7.2.3 Modified Hydrogen Atoms

The frequency of atom vibration depends on its mass. Hydrogen atoms generally have the highest vibration frequencies in the system. Such high frequencies require short time step in MD simulation and increase computational load. To limit vibration frequencies and allow a longer time step, one can apply the SHAKE algorithm to fix the length of any bond involving hydrogen atoms [175]. The SHAKE algorithm decreases the degrees of freedom in the system by introducing additional constraints between atoms. Instead, we artificially changed the mass of hydrogen atoms from 1.008 to 16.000 amu in order to preserve degree of freedom in the system following the suggestion by Bennett [176]. A larger mass of hydrogen atoms allows a longer time step in the MD algorithm. Pomes and McCammon showed that changing the hydrogen mass to 10 amu allow using a 0.01 ps time step to simulate a system which

consists of 215 TIP3P water molecules, smaller than our system [177]. Feenstra et al. change the mass of hydrogen atoms to 4 amu to increase the simulation stability of a system which contains protein and water molecules and resembles our system [178]. We set the time step as 0.001 ps, a value widely used in simulations with physical masses for all atoms, to gain higher stability in the simulation of our large system with a hemagglutinin trimer, a Fab dimer, and water molecules. As with the Einstein crystals, we exactly calculated and subtracted off the contribution of the change to the hydrogen mass to $\Delta\Delta G$. Note that the modification of hydrogen mass is independent to the reference states in the simulation, which is selected to be Einstein crystals in this project. In fact, most of the hydrogen atoms in the system are neither reactant nor product atoms. In Figure 7.1, the states with Einstein crystals and modified hydrogen atoms are states 1a, 2a, 3a, 4a, 1b, 2b, 3b, and 4b.

7.2.4 Expressions of Free Energies

Introducing two Einstein crystals and heavier hydrogen atoms changes the potential energy in the system, as well as the canonical partition functions. After modification of hydrogen atoms, the mass of atoms changed from $m_{\mathrm{r},i}$ to $m'_{\mathrm{r},i}$, from $m_{\mathrm{p},i}$ to $m'_{\mathrm{p},i}$,

or from $m_{\mathrm{x},i}$ to $m'_{\mathrm{x},i}$. Canonical partition functions of the states in Figure 7.1 are:

$$Q_3\left(n_0+n_1,V,T\right) = \frac{1}{h^{3(n_0+n_1)}} \prod_{i=1}^{n_0} \left(\frac{2\pi m_{\mathrm{x},i}}{\beta}\right)^{3/2} \prod_{i=1}^{n_1} \left(\frac{2\pi m_{\mathrm{r},i}}{\beta}\right)^{3/2} \times Z_3\left(n_0+n_1,V,T\right) \tag{7.13}$$

$$Q_{3\mathrm{a}}\left(n_0+n_1,V,T\right) = \frac{1}{h^{3(n_0+n_1)}} \prod_{i=1}^{n_0} \left(\frac{2\pi m'_{\mathrm{x},i}}{\beta}\right)^{3/2} \prod_{i=1}^{n_1} \left(\frac{2\pi m'_{\mathrm{r},i}}{\beta}\right)^{3/2} \times Z_3\left(n_0+n_1,V,T\right) \tag{7.14}$$

$$Q_{3\mathrm{b}}\left(n_0+n_1+n_2,V,T\right) = \frac{1}{h^{3(n_0+n_1)}} \prod_{i=1}^{n_0} \left(\frac{2\pi m'_{\mathrm{x},i}}{\beta}\right)^{3/2} \prod_{i=1}^{n_1} \left(\frac{2\pi m'_{\mathrm{r},i}}{\beta}\right)^{3/2} \times Z_3\left(n_0+n_1,V,T\right) \left(\frac{2\pi}{h\beta}\right)^{3n_2} \prod_{i=1}^{n_2} \left(\frac{m'_{\mathrm{p},i}}{K_i^{\mathrm{prod}}}\right)^{3/2} \tag{7.15}$$

$$Q_4\left(n_0+n_2,V,T\right) = \frac{1}{h^{3(n_0+n_2)}} \prod_{i=1}^{n_0} \left(\frac{2\pi m_{\mathrm{x},i}}{\beta}\right)^{3/2} \prod_{i=1}^{n_2} \left(\frac{2\pi m_{\mathrm{p},i}}{\beta}\right)^{3/2} \times Z_4\left(n_0+n_2,V,T\right) \tag{7.16}$$

$$Q_{4\mathrm{a}}\left(n_0+n_2,V,T\right) = \frac{1}{h^{3(n_0+n_2)}} \prod_{i=1}^{n_0} \left(\frac{2\pi m'_{\mathrm{x},i}}{\beta}\right)^{3/2} \prod_{i=1}^{n_2} \left(\frac{2\pi m'_{\mathrm{p},i}}{\beta}\right)^{3/2} \times Z_4\left(n_0+n_2,V,T\right) \tag{7.17}$$

$$Q_{4\mathrm{b}}\left(n_0+n_1+n_2,V,T\right) = \frac{1}{h^{3(n_0+n_2)}} \prod_{i=1}^{n_0} \left(\frac{2\pi m'_{\mathrm{x},i}}{\beta}\right)^{3/2} \prod_{i=1}^{n_2} \left(\frac{2\pi m'_{\mathrm{p},i}}{\beta}\right)^{3/2} \times Z_4\left(n_0+n_2,V,T\right) \left(\frac{2\pi}{h\beta}\right)^{3n_1} \prod_{i=1}^{n_1} \left(\frac{m'_{\mathrm{r},i}}{K_i^{\mathrm{reac}}}\right)^{3/2} \tag{7.18}$$

in which the states are denoted by the subscripts. Contribution of the potential energy part of the Hamiltonian to the partition function is

$$Z_3\left(n_0+n_1,V,T\right) = \int \exp\left(-\beta U_{n_0+n_1}\left(\boldsymbol{r}\right)\right) \mathrm{d}\boldsymbol{r} \tag{7.19}$$

$$Z_4\left(n_0+n_2,V,T\right) = \int \exp\left(-\beta U_{n_0+n_2}\left(\boldsymbol{r}\right)\right) \mathrm{d}\boldsymbol{r} \tag{7.20}$$

From the partition functions, free energies defined in Figure 7.1 are calculated:

$$\Delta G_{3a} = -\frac{3}{2\beta}\sum_{i=1}^{n_0}\ln\left(\frac{m'_{x,i}}{m_{x,i}}\right) - \frac{3}{2\beta}\sum_{i=1}^{n_1}\ln\left(\frac{m'_{r,i}}{m_{r,i}}\right) \tag{7.21}$$

$$\Delta G_{4a} = \frac{3}{2\beta}\sum_{i=1}^{n_0}\ln\left(\frac{m'_{x,i}}{m_{x,i}}\right) + \frac{3}{2\beta}\sum_{i=1}^{n_2}\ln\left(\frac{m'_{p,i}}{m_{p,i}}\right) \tag{7.22}$$

$$\Delta G_{3b} = -\frac{3n_2}{\beta}\ln\left(\frac{2\pi}{h\beta}\right) - \frac{3}{2\beta}\sum_{i=1}^{n_2}\ln\left(\frac{m'_{p,i}}{K_i^{\text{prod}}}\right) \tag{7.23}$$

$$\Delta G_{4b} = \frac{3n_1}{\beta}\ln\left(\frac{2\pi}{h\beta}\right) + \frac{3}{2\beta}\sum_{i=1}^{n_1}\ln\left(\frac{m'_{r,i}}{K_i^{\text{reac}}}\right) \tag{7.24}$$

$$\Delta G_{43b} = -\frac{1}{\beta}\ln\left[\frac{\prod_{i=1}^{n_1}\left(m'_{r,i}/K_i^{\text{reac}}\right)^{3/2} Z_4\left(n_0+n_2,V,T\right)}{\prod_{i=1}^{n_2}\left(m'_{p,i}/K_i^{\text{prod}}\right)^{3/2} Z_3\left(n_0+n_1,V,T\right)}\right]. \tag{7.25}$$

The free energy between state 3 and 4 is

$$\Delta G_{43} = \Delta G_{43b} - \frac{1}{\beta}\ln\frac{\left(2\pi/h\beta\right)^{3n_2}\sum_{i=1}^{n_2}\left(m_{p,i}/K_i^{\text{prod}}\right)^{3/2}}{\left(2\pi/h\beta\right)^{3n_1}\sum_{i=1}^{n_1}\left(m_{r,i}/K_i^{\text{reac}}\right)^{3/2}} = \Delta G_{43b} - \frac{1}{\beta}\ln\frac{Q_{E2}\left(n_2,V,T\right)}{Q_{E1}\left(n_1,V,T\right)} \tag{7.26}$$

in which Q_{E1} and Q_{E2} are the partition functions of the Einstein crystals for product atoms and reactant atoms, respectively. The free energy ΔG_{43b} was calculated by thermodynamic integration while ΔG_{43} was used to calculate the free energy difference of one substitution. Note that the correction term between ΔG_{43b} and ΔG_{43} is independent of the masses of atoms. Canonical partition functions as well as free energies of the state 1, 1a, 1b, 2, 2a, and 2b are calculated in a similar way.

7.2.5 Implementation of Free Energy Calculation Algorithm

The above discussion is the theoretical basis for the implementation of our free energy calculation algorithm. The free energy calculation protocol consists of four steps. First, we built the dual topology with reactant and product atoms in the amino acid substitution site in separated antibody and hemagglutinin or antibody-hemagglutinin

complex. We then solvated the protein system and modified the mass of hydrogen atoms. Second, two Einstein crystals were introduced as the reference states for the reactant and product atoms, respectively. Third, the MD simulation was run at 64 windows. The thermodynamic integration algorithm obtained the free energy values ΔG_{21} for separated antibody and hemagglutinin or ΔG_{43} for antibody-hemagglutinin complex, as in Figure 7.1. This step gave the $\Delta\Delta G$ value. Fourth, we calculated the error bar of the $\Delta\Delta G$ value obtained in the last step. The technical details of these four steps are illustrated in the text below. Also described are the verification of the free energy calculation protocol, the software and hardware information, and the CPU hours consumed by the protocol.

The hemagglutinin trimer of H3N2 virus strain A/Aichi/2/1968 with bound dimer antibody HC63 (PDB code: 1KEN) was used in our calculation. For each amino acid substitution, we built the dual topology with side chains of both amino acids prior to the simulation. Reactant and product atoms were defined as the side chains in the original and substituting amino acid, respectively. All the covalent and non-bonded interactions between reactant and product atoms were removed. The protein was in an explicit water box with periodic boundary condition. The mass of hydrogen atoms was changed from 1.008 to 16.000 amu.

All the simulations were performed by CHARMM c33b2 with CHARMM22 force field [152]. We first fixed the positions of hemagglutinin trimer except reactant atoms and minimized the system with 200 steps of steepest descent (SD) algorithm and 5000 steps of adopted basis Newton-Raphson (ABNR) algorithm. We ran a 5 ps MD simulation of the system, the trajectory of which gave the spatial fluctuation $\left\langle (\delta \boldsymbol{r}_i)^2 \right\rangle$ of each reactant atom. Then we fixed reactant atoms, released product atoms, and ran a 5 ps MD simulation to obtain the spatial fluctuation of each product

atom. Final positions of both reactant and product were adopted as the reference positions of the corresponding Einstein crystal. The force constant K_i of each atom in Einstein crystals was obtained from $\left\langle (\delta \boldsymbol{r}_i)^2 \right\rangle$ by equation 7.11 and equation 7.12. With modified hydrogen atoms and two Einstein crystals as the reference states of reactant and product atoms, state 1b, 2b, 3b, and 4b in Figure 7.1 were generated for thermodynamic integration.

In thermodynamic integration, MD simulations were run at 64 windows with distinct λ. In each window, pressure of the system was first calibrated with a 10 ps MD simulation in an isothermal-isobaric (NPT) ensemble. The duration of 10 ps is appropriate because it is long enough to equilibrate the pressure and short enough to prevent the protein from drifting away from the original location. We fixed coordinates of the residues and water molecules except for those within 15 Å from the three alpha carbons. Then we removed amino acid residues and water molecules other than those within 27.5 Å from the three alpha carbons of substituted residues in the hemagglutinin trimer to reduce the system size, because the fixed atoms are not included in the topology of movable atoms and the cutoff of the non-bonded forces is 12 Å. The Ewald sum was used to calculate charge interactions. Note that this substantial reduction of the system relies on the assumption that the free energy change due to the amino acid substitution is mostly affected by atoms near the binding site after the system reaches equilibrium. This assumption is based on two facts: the conformations of hemagglutinin and antibody are stable once the system reaches equilibrium, and all the removed or fixed atoms have invariant interactions with the substituting amino acid residues. The stable protein conformation means amino acid residues far away from the substituting residue do not move during the amino acid substitution process. In the CHARMM22 force field used in this project, the cutoff of non-bonded force is

12 Å and less than the 15 Å threshold for system reduction. The system reduction does not directly affect the force on the substituted residue because of absence of the long-range non-bonded force between the substituted residue and atoms removed from the system. This system reduction method was also applied to compute binding free energy of subtilisin [179], of tripsin [163], and of Src SH2 domain [164]. Robust results were obtained in all of these applications. Generally, this system reduction strategy can produce reliable result if the reduced system contains the residues and molecules critical to the binding process [163]. We note that the system reduction method could be a limitation of the free energy calculation model. The fixing of amino acid residues and water molecules described in section 2.5 substantially reduced the CPU time needed, but is an approximation to the real system containing the whole proteins. This limitation reflects the tradeoff between model accuracy and required computational resource. In the canonical ensemble, the new system was equilibrated for 200 ps and simulated for another 900 ps as the data production phase. The integrand of thermodynamic integration is the ensemble average of the sampled trajectory $\langle \partial U(\boldsymbol{r},\lambda)/\partial\lambda \rangle_\lambda = \langle U_{\text{ein,reac}}(\boldsymbol{r}) - U_{\text{reac}}(\boldsymbol{r}) - U_{\text{ein,prod}}(\boldsymbol{r}) + U_{\text{prod}}(\boldsymbol{r}) \rangle_\lambda$. The free energy ΔG_{21} and ΔG_{43} between the real states was calculated by adding a correction term of the Einstein crystals in equation 7.26. Finally, the difference of antibody binding free energy is $\Delta\Delta G = \Delta G_{43} - \Delta G_{21}$.

Error bars of $\Delta\Delta G$ are also given. The convergence behavior of the simulation was analyzed using the block average method developed by Flyvbjerg and Petersen [180]. As mentioned above, the MD simulation for either the unbound hemagglutinin or the hemagglutinin-antibody complex contains 64 windows with distinct λ. The 900 ps data production phase contains 9×10^5 simulation steps. The values $A = U_{\text{prod}}(r) - U_{\text{reac}}(r)$, as in equation equation 7.3, computed in consecutive simulation

steps were grouped into bins, and consecutive bins were merged progressively. The quantity $\sigma^2(A)/(n-1)$, in which $\sigma^2(A)$ is the variance of the average of each bin $A_1, A_2, \ldots, A_n$ and n is the number of bins, increases with the bin size and reaches a plateau when the bin size is 1×10^4 steps. We fixed the bin size to 1×10^4 steps and estimate the variance of ensemble average $\langle A \rangle$ as $\sigma^2(A)/(n-1)$, following Flyvbjerg and Petersen's method [180].

This protocol, without the Einstein crystal contribution, was verified by recalculating published free energy differences of amino acid substitution T131I [75]. Without the Einstein crystal contribution, our protocol gave the $\Delta\Delta G = 5.69 \pm 0.07$ kcal/mol, compared to the $\Delta\Delta G = 5.20 \pm 0.94$ kcal/mol in the published work [75]. Theoretically exact results presented here include the Einstein crystal contribution. We note that the theoretically exact $\Delta\Delta G$ for T131I, including the Einstein crystal contribution, is 3.71 ± 0.07 kcal/mol.

The simulation was performed using CHARMM22 force field at three clusters: tg-steele.purdue.teragrid.org (Intel Xeon E5410, 2.33 GHz), sugar.rice.edu (Intel Xeon E5440, 2.83 GHz), and biou.rice.edu (IBM POWER7, 3.55 GHz), as well as at the condor pool tg-condor.rcac.purdue.edu at Purdue University. Simulation of each substitution took approximately 7.5 thousand CPU hours on average, and so this work consumed about three million CPU hours.

7.3 Results

7.3.1 Free Energy Landscape

For each of the 21 amino acid sites in epitope B, we substituted from alanine to each one of the 19 other amino acids, in which we used the neutral histidine (CHARMM

code: Hse) as the model of histidine. The free energy difference and standard error of each substitution were calculated by the MD simulation (see Materials and Methods). The wildtype amino acid in each site of epitope B was extracted from the hemagglutinin sequence of the H3N2 strain A/Aichi/2/1968. The free energy difference and standard error of the substitution from the wildtype amino acid in each site were then calculated from the values for the change from the wildtype amino acid to alanine and from alanine to the new amino acid. The values are listed in Table 7.1.

Table 7.1 : Summary of the calculated free energy differences $\Delta\Delta G$ in each amino acid site in epitope B from the wildtype amino acid to all 20 amino acids. The standard errors are also listed. The free energy difference and its standard error of the substitution from the wildtype amino acid to itself are both zero. The units of free energy differences and their standard errors are kcal/mol.

Positions	128	129	155	156	157	158	159
Ala	-13.12 ± 0.27	3.33 ± 0.29	2.78 ± 0.20	1.19 ± 0.33	2.48 ± 0.21	4.27 ± 0.31	5.18 ± 0.21
Arg	22.57 ± 0.46	2.31 ± 0.45	16.98 ± 0.37	0.08 ± 0.50	-4.19 ± 0.44	-1.61 ± 0.48	7.07 ± 0.42
Asn	-4.80 ± 0.36	5.83 ± 0.42	-7.83 ± 0.30	10.72 ± 0.40	5.64 ± 0.34	3.41 ± 0.42	10.97 ± 0.35
Asp	4.52 ± 0.38	19.12 ± 0.42	16.28 ± 0.32	11.06 ± 0.42	9.95 ± 0.37	18.37 ± 0.40	15.34 ± 0.36
Cys	-11.83 ± 0.34	12.64 ± 0.37	-2.37 ± 0.30	5.32 ± 0.38	-2.72 ± 0.29	-7.88 ± 0.40	7.92 ± 0.32
Gln	-12.37 ± 0.40	7.34 ± 0.42	-4.29 ± 0.36	13.14 ± 0.41	-0.45 ± 0.36	11.47 ± 0.43	6.54 ± 0.40
Glu	11.15 ± 0.38	10.50 ± 0.42	17.77 ± 0.34	26.54 ± 0.43	4.68 ± 0.36	8.58 ± 0.48	5.19 ± 0.39
Gly	-9.93 ± 0.39	0.00 ± 0.00	17.00 ± 0.34	0.11 ± 0.44	0.21 ± 0.36	0.00 ± 0.00	-4.19 ± 0.41
Hse	4.43 ± 0.42	0.15 ± 0.43	2.47 ± 0.36	-6.89 ± 0.43	12.18 ± 0.38	5.54 ± 0.46	1.06 ± 0.39
Ile	-16.03 ± 0.41	0.54 ± 0.40	1.55 ± 0.33	8.33 ± 0.42	11.22 ± 0.37	8.09 ± 0.43	18.96 ± 0.39
Leu	-23.58 ± 0.41	-4.27 ± 0.43	-8.92 ± 0.33	2.64 ± 0.45	-6.26 ± 0.39	1.61 ± 0.45	4.08 ± 0.38
Lys	3.57 ± 0.45	11.18 ± 0.46	14.58 ± 0.37	0.00 ± 0.00	6.24 ± 0.40	-1.60 ± 0.48	5.39 ± 0.46
Met	-13.38 ± 0.39	-2.59 ± 0.39	1.23 ± 0.35	10.11 ± 0.43	16.15 ± 0.36	14.49 ± 0.44	-6.38 ± 0.37
Phe	-10.21 ± 0.45	6.12 ± 0.43	9.39 ± 0.35	0.30 ± 0.45	10.28 ± 0.40	5.17 ± 0.48	12.33 ± 0.42
Pro	-9.36 ± 0.36	-2.43 ± 0.42	-1.86 ± 0.31	2.32 ± 0.43	5.69 ± 0.30	17.09 ± 0.40	6.08 ± 0.36
Ser	-14.55 ± 0.34	3.36 ± 0.37	-1.09 ± 0.29	-1.45 ± 0.38	0.00 ± 0.00	2.76 ± 0.39	0.00 ± 0.00
Thr	0.00 ± 0.00	7.35 ± 0.36	0.00 ± 0.00	-1.08 ± 0.41	6.34 ± 0.32	8.36 ± 0.41	15.32 ± 0.32
Trp	9.82 ± 0.47	4.81 ± 0.47	19.84 ± 0.43	23.26 ± 0.48	16.14 ± 0.45	3.52 ± 0.62	-1.35 ± 0.45
Tyr	-14.83 ± 0.43	2.72 ± 0.42	7.25 ± 0.36	-2.18 ± 0.46	-8.37 ± 0.44	18.42 ± 0.51	5.95 ± 0.43
Val	-19.13 ± 0.37	3.56 ± 0.38	8.57 ± 0.31	-3.01 ± 0.39	7.63 ± 0.32	3.77 ± 0.42	6.45 ± 0.32

Positions	160	163	165	186	187	188	189
Ala	4.16 ± 0.22	−0.24 ± 0.22	4.15 ± 0.24	−3.19 ± 0.19	−4.03 ± 0.23	3.45 ± 0.25	−9.01 ± 0.28
Arg	9.70 ± 0.44	5.97 ± 0.39	14.58 ± 0.41	21.01 ± 0.38	8.12 ± 0.42	−0.06 ± 0.45	−0.39 ± 0.48
Asn	2.07 ± 0.34	−2.32 ± 0.32	0.00 ± 0.00	4.67 ± 0.30	−10.07 ± 0.34	0.00 ± 0.00	−3.18 ± 0.37
Asp	13.50 ± 0.32	12.64 ± 0.32	25.01 ± 0.31	24.54 ± 0.28	7.78 ± 0.35	19.77 ± 0.37	6.77 ± 0.35
Cys	15.82 ± 0.31	1.84 ± 0.30	1.93 ± 0.29	−2.30 ± 0.25	−11.09 ± 0.32	4.07 ± 0.34	6.23 ± 0.33
Gln	3.04 ± 0.39	−8.29 ± 0.35	4.27 ± 0.36	5.16 ± 0.33	−2.87 ± 0.37	12.36 ± 0.39	0.00 ± 0.00
Glu	15.48 ± 0.36	2.17 ± 0.35	15.74 ± 0.34	33.29 ± 0.31	14.41 ± 0.35	10.10 ± 0.37	12.16 ± 0.39
Gly	1.22 ± 0.39	−5.83 ± 0.38	9.11 ± 0.37	0.13 ± 0.27	−0.60 ± 0.30	−5.06 ± 0.32	−5.69 ± 0.32
Hse	0.52 ± 0.38	6.31 ± 0.33	7.44 ± 0.33	18.15 ± 0.30	3.69 ± 0.39	−1.95 ± 0.40	−8.53 ± 0.40
Ile	1.51 ± 0.34	10.62 ± 0.36	3.85 ± 0.33	−1.85 ± 0.30	−2.51 ± 0.33	−4.77 ± 0.37	3.65 ± 0.37
Leu	−1.39 ± 0.40	3.85 ± 0.35	−9.20 ± 0.37	1.07 ± 0.30	−0.40 ± 0.38	−1.30 ± 0.37	−6.91 ± 0.39
Lys	5.91 ± 0.44	10.37 ± 0.38	1.93 ± 0.41	−1.15 ± 0.39	24.91 ± 0.41	8.42 ± 0.44	9.48 ± 0.64
Met	10.78 ± 0.38	7.22 ± 0.35	1.63 ± 0.36	13.06 ± 0.33	−5.11 ± 0.36	6.97 ± 0.38	6.86 ± 0.40
Phe	7.90 ± 0.41	−0.86 ± 0.36	13.87 ± 0.38	6.94 ± 0.33	−7.23 ± 0.39	2.05 ± 0.39	4.37 ± 0.43
Pro	4.51 ± 0.32	12.50 ± 0.34	18.96 ± 0.33	11.82 ± 0.29	10.69 ± 0.32	−10.24 ± 0.35	−8.98 ± 0.36
Ser	7.13 ± 0.29	9.07 ± 0.30	−0.92 ± 0.28	0.00 ± 0.00	−4.88 ± 0.31	8.09 ± 0.33	−5.09 ± 0.34
Thr	0.00 ± 0.00	9.18 ± 0.30	10.35 ± 0.31	−14.79 ± 0.27	0.00 ± 0.00	3.53 ± 0.38	9.30 ± 0.35
Trp	0.86 ± 0.44	12.34 ± 0.35	19.02 ± 0.43	−7.69 ± 0.38	−11.04 ± 0.48	7.20 ± 0.40	−9.19 ± 0.45
Tyr	−5.43 ± 0.39	1.06 ± 0.34	14.76 ± 0.37	11.90 ± 0.33	5.29 ± 0.42	1.57 ± 0.40	4.81 ± 0.41
Val	7.99 ± 0.34	0.00 ± 0.00	9.79 ± 0.32	2.97 ± 0.29	3.08 ± 0.33	3.73 ± 0.34	−7.89 ± 0.36

Positions	190	192	193	194	196	197	198
Ala	−18.12 ± 0.24	−0.86 ± 0.23	−5.20 ± 0.20	2.37 ± 0.23	5.95 ± 0.23	−2.40 ± 0.29	0.00 ± 0.00
Arg	4.97 ± 0.41	23.07 ± 0.44	32.33 ± 0.41	−13.66 ± 0.37	−25.38 ± 0.44	−17.94 ± 0.47	3.99 ± 0.37
Asn	−16.44 ± 0.30	−2.56 ± 0.32	8.24 ± 0.30	−3.81 ± 0.31	13.27 ± 0.36	−6.58 ± 0.38	0.05 ± 0.28
Asp	18.75 ± 0.32	2.92 ± 0.32	15.29 ± 0.29	26.72 ± 0.35	9.25 ± 0.34	5.58 ± 0.39	5.17 ± 0.24
Cys	−20.36 ± 0.32	−1.45 ± 0.32	−9.79 ± 0.26	1.91 ± 0.30	1.30 ± 0.31	6.70 ± 0.36	5.91 ± 0.22
Gln	−17.37 ± 0.37	−6.00 ± 0.37	4.87 ± 0.34	−0.83 ± 0.32	7.68 ± 0.36	0.00 ± 0.00	1.41 ± 0.31
Glu	0.00 ± 0.00	1.18 ± 0.35	45.40 ± 0.34	38.35 ± 0.33	3.60 ± 0.36	11.34 ± 0.41	2.37 ± 0.30
Gly	−17.09 ± 0.29	−13.46 ± 0.30	−13.89 ± 0.27	−18.59 ± 0.30	8.08 ± 0.31	4.11 ± 0.36	3.65 ± 0.28
Hse	−26.26 ± 0.35	−0.96 ± 0.38	−0.96 ± 0.35	9.95 ± 0.34	18.42 ± 0.37	−2.62 ± 0.40	−3.27 ± 0.35
Ile	−16.45 ± 0.37	−5.57 ± 0.37	−3.80 ± 0.31	−6.91 ± 0.32	0.77 ± 0.34	1.23 ± 0.41	0.01 ± 0.32
Leu	−17.27 ± 0.36	−7.97 ± 0.37	10.76 ± 0.34	0.00 ± 0.00	10.07 ± 0.39	−0.03 ± 0.40	−11.18 ± 0.29
Lys	−9.33 ± 0.38	5.67 ± 0.42	39.36 ± 0.39	−16.67 ± 0.38	0.49 ± 0.40	−16.50 ± 0.47	1.98 ± 0.37
Met	−26.63 ± 0.34	6.82 ± 0.36	−2.91 ± 0.32	7.75 ± 0.35	4.08 ± 0.37	−7.79 ± 0.40	15.57 ± 0.32
Phe	−31.89 ± 0.39	1.56 ± 0.40	16.46 ± 0.59	2.78 ± 0.34	−1.99 ± 0.37	1.05 ± 0.44	8.73 ± 0.34
Pro	−17.85 ± 0.33	−2.28 ± 0.33	9.84 ± 0.31	8.01 ± 0.31	15.42 ± 0.35	−5.34 ± 0.40	0.70 ± 0.29
Ser	−14.75 ± 0.31	−7.79 ± 0.30	0.00 ± 0.00	6.62 ± 0.29	6.91 ± 0.29	1.97 ± 0.36	−2.40 ± 0.22
Thr	−4.17 ± 0.32	0.00 ± 0.00	−2.04 ± 0.27	12.40 ± 0.31	7.81 ± 0.33	−7.91 ± 0.36	6.79 ± 0.24
Trp	−22.93 ± 0.39	2.31 ± 0.44	17.92 ± 0.42	−1.30 ± 0.40	8.17 ± 0.43	−7.73 ± 0.44	−7.23 ± 0.38
Tyr	−13.82 ± 0.38	7.63 ± 0.42	16.16 ± 0.38	9.73 ± 0.36	2.92 ± 0.40	6.10 ± 0.44	−4.82 ± 0.32
Val	−9.12 ± 0.31	−6.80 ± 0.32	−6.92 ± 0.30	2.59 ± 0.29	0.00 ± 0.00	4.16 ± 0.39	−4.22 ± 0.24

As described in equation 7.26, each $\Delta\Delta G$ value listed in Table 7.1 contains the contribution of two Einstein crystals. The contribution of Einstein crystals to the final

$\Delta\Delta G$ values was calculated for each of the 399 amino acid substitutions in epitope B. The average fraction of the contribution of Einstein crystals in the calculated $\Delta\Delta G$ values is 44%. The contribution of Einstein crystals is far greater than that of the statistical error of our free energy calculation in Table 7.1, which is 4.5% on average. Thus, the Einstein crystal contribution is both theoretically exact and practically important. In 371 of the 399 substitutions, the absolute values of the contribution of Einstein crystals is greater than 1.96 standard errors of the final $\Delta\Delta G$ values. That is, the contribution of Einstein crystals is significant with $p < 0.05$ in 93.0% of all the amino acid substitutions. Consequently, it is essential to incorporate Einstein crystals in the free energy calculation to eliminate the error caused by the methods that neglect the unknown effect of the translational entropy of the free atoms in thermodynamic integration. The contribution of the translational entropy of ideal gas-like atoms ($\lambda = 0$ or $\lambda = 1$) needs to be either calculated or removed by a theoretically exact method to perform an exact free energy calculation.

The obtained $\Delta\Delta G$ values allow us to analyze the character of each of the 20 amino acids. We first averaged over all the 21 amino acid sites in epitope B the $\Delta\Delta G$ value caused by the single substitutions from alanine to the other amino acids. The averaged $\Delta\Delta G$ values are listed in Table 7.2. The largest $\Delta\Delta G$ are caused by the negatively charged amino acids (Glu, Asp) and the positively charged amino acids (Arg, Lys), indicating that introduction of charged amino acids in the dominant epitope decreases the binding affinity between antibody and hemagglutinin. Note that amino acid substitutions that change the charge of hemagglutinin significantly affect the calculated free energy values [181, 182, 183]. The issue of how to best calculate free energy differences when charge changes has been debated over the years. In the present paper, we are using the standard Ewald approach with explicit solvent. We

note that the evolutionary history of H3 hemagglutinin since 1968 shows an increasing trend of the number of charged amino acids in epitope B [132], which agrees with the results that introduction of charged amino facilitates virus evasion from antibody, as illustrated in Table 7.2. The result that introduction of charged amino acid on average increases $\Delta\Delta G$ is not an artifact, is supported by data from the influenza evolution, and is expected on the basis that charge is hydrophilic. In addition to the charge, the rank of free energy differences also largely correlated to the size of amino acid. By the definition used by RasMol [13], the 16 uncharged amino acids are tagged as hydrophobic (Ala, Gly, Ile, Leu, Met, Phe, Pro, Trp, Tyr, Val), large (Gln, Hse, Ile, Leu, Met, Phe, Trp, Tyr), medium (Asn, Cys, Pro, Thr, Val), and small (Ala, Gly, Ser), as shown in Table 7.2. The ranks of small amino acids are lower than those of medium amino acids ($p = 0.036$, Wilcoxon rank-sum test) and those of large amino acids ($p = 0.085$, Wilcoxon rank-sum test). In contrast, the hydrophobicity of the uncharged amino acids is largely uncorrelated to their ranks by $\Delta\Delta G$. As a result, charged amino acids in the dominant epitope are essential to the immune evasion while the virus escape substitution among small amino acids have minimal effect.

Epitope B comprises 21 amino acid sites in the top of the hemagglutinin trimer. Taking the probability for one substituting amino acid to exist at each site to be proportional to the relative frequency of this amino acid in H3 hemagglutinin, the weighted average free energy difference in each of the 21 sites was calculated. The relative frequencies of 20 amino acids were obtained from 6896 H3 hemagglutinin sequences deposited between 1968 and 2009 in the NCBI database [184] and listed in Table 7.2. Also using the $\Delta\Delta G$ values in Table 7.1, we calculated and tabulated in Table 7.3 for each site i the value of $\langle\Delta\Delta G\rangle_i$, which is the average $\Delta\Delta G$ weighted by the probability for each different amino acid to be introduced, where probability

Table 7.2 : The rank of the average binding free energy difference of the single substitution from alanine to another amino acid over all the 21 amino acid sites in epitope B of hemagglutinin trimer. The rank correlates with the charge and the size of amino acid, and it is relatively uncorrelated to the hydrophobicity. Here we applied classifications of RasMol for the biochemical properties of the 20 amino acids [13]. The relative frequencies of 20 amino acids were counted from the H3 sequences in NCBI database from 1968 to 2009.

Rank	Amino Acid	$\Delta\Delta G$ (kcal/mol)	Charged	Hydrophobic	Large	Medium	Small	Relative frequency
1	Glu	14.612 ± 0.061	×		×			0.029
2	Asp	14.533 ± 0.055	×			×		0.051
3	Arg	6.018 ± 0.078	×		×			0.052
4	Lys	5.766 ± 0.078	×		×			0.057
5	Trp	4.458 ± 0.081		×	×			0.016
6	Tyr	3.984 ± 0.071		×	×			0.035
7	Thr	3.981 ± 0.050				×		0.078
8	Pro	3.912 ± 0.054		×		×		0.060
9	Met	3.562 ± 0.062		×	×			0.009
10	Phe	3.522 ± 0.073		×	×			0.030
11	Hse	2.654 ± 0.064			×			0.020
12	Gln	1.985 ± 0.063			×			0.042
13	Ile	1.396 ± 0.060		×	×			0.070
14	Asn	1.150 ± 0.054				×		0.085
15	Val	1.147 ± 0.051		×		×		0.055
16	Cys	0.888 ± 0.046				×		0.028
17	Ser	0.469 ± 0.044					×	0.096
(18)	(Ala)	(0.000 ± 0.000)		×			×	0.046
19	Gly	-1.612 ± 0.055		×			×	0.070
20	Leu	-2.273 ± 0.064		×	×			0.071

is proportional to the relative frequencies of 20 amino acids counted from the H3 sequences in NCBI database from 1968 to 2009.

As shown in Table 7.3, there is obvious variation among the expected free energy differences $\langle \Delta\Delta G \rangle_i$ caused by single substitutions at amino acid site i of epitope B. This variation is partly due to the wildtype amino acids in the sites. For instance, the wildtype amino acid in site 190 is Glu that has the highest rank in Table 7.2. As shown in Table 7.3, any amino acid substitution in site 190 tends to have a negative $\Delta\Delta G$. Another cause of variation in $\langle \Delta\Delta G \rangle_i$ is that distinct sites affect differently the antibody binding process. Epitope B of the wildtype A/Aichi/2/1968 hemagglutinin sequence contains five sites with threonine: 128, 155, 160, 187, and 192. The mathematical expectancies $\langle \Delta\Delta G \rangle_i$ in these five sites are -7.746, 4.471, 4.956, 1.182, and -1.737 kcal/mol, respectively. Therefore, each site in epitope B has a specific effect on the virus escape substitution. A random substitution in epitope B affects the antibody binding free energy differently depending on the site and the substituting amino acids.

The variation of $\langle \Delta\Delta G \rangle_i$ is also reflected by the tertiary structure of the epitope B bound by the antibody. By looking into the structure of epitope B shown in Figure 7.2. Epitope B resides in two protruding loops from amino acid site 128 to 129, and from site 155 to 165, respectively, and in a α-helix from site 186 to 198. Site 128 has a negative average free energy difference $\langle \Delta\Delta G \rangle_{128} = -7.746 \pm 0.098$ kcal/mol. All the other sites in these two loops show a positive $\langle \Delta\Delta G \rangle_i$ value of a random substitution, with the minimum $\langle \Delta\Delta G \rangle_{157} = 3.944 \pm 0.090$ kcal/mol in site 157. The α-helix is located between hemagglutinin and antibody. In the α-helix, the sites facing towards the antibody usually present large positive $\langle \Delta\Delta G \rangle_i$ values such as site 193 and 196, while the sites facing towards the hemagglutinin show lower $\langle \Delta\Delta G \rangle_i$

Table 7.3 : The rank of the average free energy difference $\langle \Delta\Delta G \rangle_i$ generated by a substitution in each amino acid site i of epitope B.

Rank	Site	$\langle \Delta\Delta G \rangle_i$ (kcal/mol)
1	193	8.074 ± 0.081
2	159	7.792 ± 0.094
3	165	7.741 ± 0.086
4	158	6.128 ± 0.108
5	196	5.444 ± 0.088
6	160	4.956 ± 0.090
7	186	4.754 ± 0.076
8	163	4.722 ± 0.085
9	129	4.690 ± 0.103
10	155	4.471 ± 0.081
11	156	4.029 ± 0.106
12	157	3.944 ± 0.090
13	188	2.945 ± 0.092
14	194	1.886 ± 0.080
15	187	1.182 ± 0.087
16	198	0.531 ± 0.072
17	189	-0.631 ± 0.098
18	192	-1.737 ± 0.087
19	197	-1.967 ± 0.099
20	128	-7.746 ± 0.098
21	190	-12.666 ± 0.084

such as site 189, 192, and 197. Thus in the one dimensional sequence from site 186 to 198, the $\langle \Delta\Delta G \rangle_i$ values oscillate with peaks and valleys corresponding to the sites in the α-helix facing alternatingly to the antibody and hemagglutinin. Consequently, the variation of the expected free energy changes in distinct sites depends on the structure of the hemagglutinin-antibody complex.

7.3.2 Historical Substitutions in Epitope B

The simulation results are supported in two aspects by amino acid sequence data of H3 hemagglutinin collected since 1968. These historical sequences are downloaded from the NCBI Influenza Virus Resource [185] and aligned. First, Pan et al. analyzed the number of charged amino acid in epitope B of H3 hemagglutinin in each year since 1968, and found an increasing trend of charged amino acids [132]. This finding supports the results that amino acid substitution introducing charged residues on average facilitates virus escape from antibody, as illustrated in Table 7.2. Second, amino acid substitutions in epitope B between 1968 and 1975 also verified the free energy calculation, as shown below.

With the knowledge of the free energy landscape of the single substitutions, we are able to recognize favorable single substitutions in epitope B. Substitutions with large positive $\Delta\Delta G$ values enable the virus to evade the immune pressure and increase the virus fitness. Favorable substitutions grow in the virus population. Selection for substitutions with large $\Delta\Delta G$ is part of the evolutionary strategy of the virus. The results of free energy calculation can also explain the substituted virus strains collected in history.

We analyzed the hemagglutinin sequence information of H3N2 strains evolving from the A/Aichi/2/1968 strains. H3 hemagglutinin circulating from 1968 to 1971

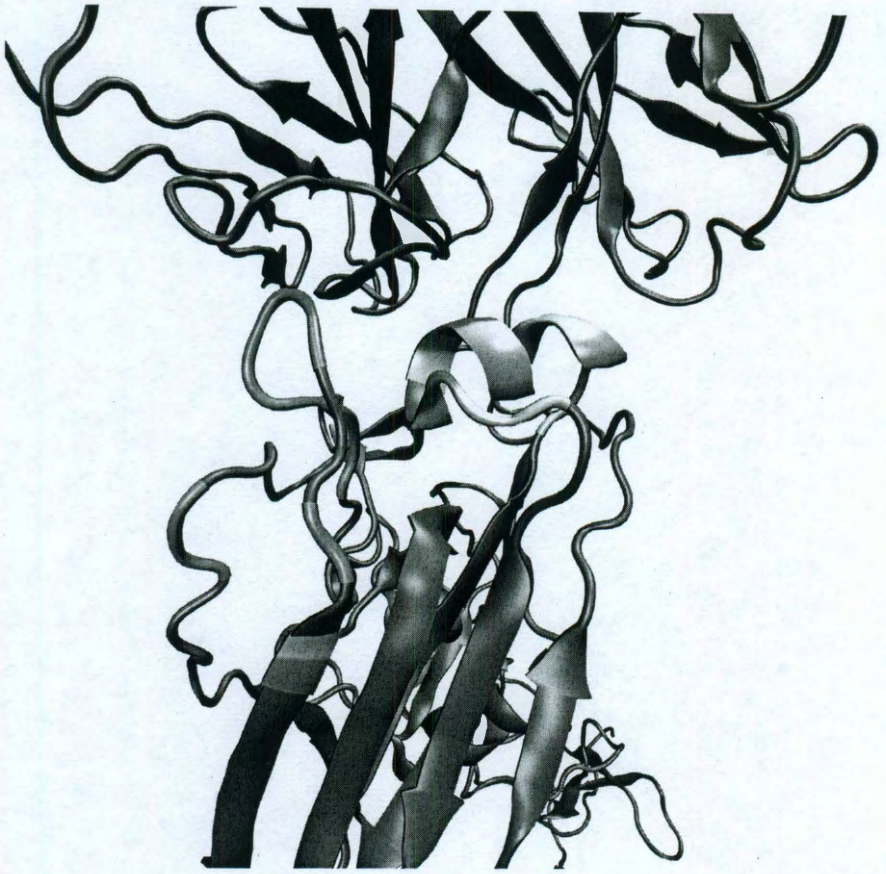


Figure 7.2 : The tertiary structure of the interface between the HA1 domain of H3 hemagglutinin monomer A/Aichi/2/1968 (bottom) and the antibody HC63 (top) (PDB code: 1KEN). Water molecules are not shown. Epitope B of the HA1 domain is located in two loops and one α-helix with the color scale modulated according to the expected free energy difference $\langle \Delta\Delta G \rangle_i$ of each site i in epitope B. The color scale ranges from red for the most negative $\langle \Delta\Delta G \rangle_i$ values to blue for the most positive $\langle \Delta\Delta G \rangle_i$ values. The sites i in epitope B with $\langle \Delta\Delta G \rangle_i$ near zero are colored white. The region outside epitope B is colored gray. The red site 128 is far from the antibody binding region and the red site 190 possessed the original amino acid Glu, which is a charged amino acid. It may explain why these two sites show negative $\langle \Delta\Delta G \rangle_i$ with large absolute values.

was mainly in the HK68 antigenic cluster while those circulating from 1972 to 1975 were mainly in the EN72 antigenic cluster [30]. Table 7.4 shows that in the dominant epitope B, there were 17 substitutions occurred in 12 sites collected between 1968 to 1975 [4], which contributed to the immune evasion and corresponding virus evolution from the HK68 cluster to the EN72 cluster. Also listed in Table 7.4 are the free energy differences of these historical substitutions. The 17 substituting amino acids have significantly higher ranks compared to the corresponding wildtype amino acids ($p = 0.0044$, Wilcoxon signed-rank test). This significant difference is expected because 15 of 17 substituting amino acids have ranks between 1 and 10, while 10 of 12 wildtype amino acids in the substituted site have ranks between 11 and 20. In all the 21 sites in epitope B, 15 of 21 wildtype amino acids have ranks between 11 and 20. Additionally, the $\Delta\Delta G$ values of these 17 substitutions listed in Table 7.4 are greater than the expected free energy differences $\langle\Delta\Delta G\rangle_i$ in Table 7.3 of random substitutions in the 12 substituted sites ($p = 0.013$, Wilcoxon signed-rank test).

We also looked into the historical escape substitutions in epitope B evading the immune pressure of the vaccine strains. For each influenza season, the amino acids in the administered vaccine strain were defined as the wildtype ones and those in the dominant circulating strain as the substituting amino acids. In each of the 19 seasons in which H3N2 virus was the dominant subtype from 1971 to 2004, the substitutions in epitope B were located [2] and their $\Delta\Delta G$ values were obtained from Table 7.1. As shown in Table 7.5, escape substitutions in epitope B as of 1973 mostly had positive $\Delta\Delta G$ and generated substituting amino acids with increased rank ($p = 0.047$, Wilcoxon signed-rank test). Such tendency to introduce amino acids with higher ranks was not observed after 1973: the ranks of wildtype and substituting amino acids after 1973 present little significant difference ($p = 0.28$, Wilcoxon signed-rank test). The

Table 7.4 : Substitutions occurred in epitope B of the hemagglutinin A/Aichi/2/1968 (H3N2) as of 1975. Also listed are the time when the substitutions were observed, and the free energy differences with standard errors. In each site of epitope B, all the 20 amino acid were sorted in the descending order by the free energy differences introduced by a substitution from the wildtype amino acid to 20 amino acids. The ranks of the substituting amino acid and the wildtype amino acid in each substituted site are listed in the column Rank (substituting) and Rank (WT), respectively.

Substitution	Year	$\Delta\Delta G$ (kcal/mol)	Rank (substituting)	Rank (WT)
T128N	1971	-4.796 ± 0.361	8	7
T128I	1975	-16.026 ± 0.412	18	7
G129E	1970, 1972	10.500 ± 0.415	4	17
T155Y	1972–1973, fixed in 1973	7.254 ± 0.358	9	14
G158E	1971–1972	8.584 ± 0.479	6	17
S159N	1971, 1974–1975	10.969 ± 0.352	5	17
S159C	1972	7.923 ± 0.324	6	17
S159R	1972	7.065 ± 0.424	7	17
T160A	1973	4.160 ± 0.217	11	18
S186N	1975	4.673 ± 0.298	10	14
N188D	1971–1973, fixed in 1973	19.767 ± 0.367	1	14
Q189K	1975	9.484 ± 0.640	2	10
E190V	1972	-9.115 ± 0.310	5	3
E190D	1975	18.752 ± 0.324	1	3
S193N	1972–1975	8.239 ± 0.301	10	12
S193D	1975	15.285 ± 0.294	7	12
A198T	1972	6.793 ± 0.236	3	14

hemagglutinin of A/Aichi/2/1968 used in the free energy calculating is in the HK68 antigenic cluster. Perhaps after the virus evolved into the next EN72 cluster, change in the virus antigenic character stimulates the immune system to produce new types of antibody other than the HC63 antibody used in the calculation. A different binding antibody changes the free energy landscape of the substitutions in epitope B. Thus the application of the present free energy landscape should be limited within the HK68 and EN72 clusters. Free energy differences of substitutions in the EN72 cluster would need to be calculated using the updated antibody crystal structure.

7.4 Discussion

7.4.1 Fitness of the Virus Strains

The free energy landscape shown in Table 7.1 gives the change of the antibody binding affinity, $K_1/K_0 = \exp(-\Delta\Delta G/RT)$, induced by each possible substitution in epitope B of the wildtype hemagglutinin. The majority of the substitutions lead to positive $\Delta\Delta G$, and yield a reduced binding affinity K_1 that is smaller than the binding affinity of the original mature antibody K_0. Decreased antibody binding constant grants the virus a higher chance of evading the immune pressure and infecting host cells. We propose that virus fitness is positively correlated to the free energy difference $\Delta\Delta G$. The other factor affecting virus fitness is the capability of the hemagglutinin to maintain the normal biochemical functions, such as virus entry. Most sites in epitope B changed amino acid identities during 1968 to 2005 as the H3N2 virus kept circulating [4]. We therefore postulate that the substitutions in epitope B do not greatly interfere with the biochemical function of hemagglutinin, and virus fitness is dominantly determined by the free energy difference resulted from substitutions in

Table 7.5 : Substitutions occurred in epitope B of H3 hemagglutinin between the vaccine strain and the dominant circulating strain in each season in which the H3N2 subtype was dominant. The free energy difference with standard error of each substitution is obtained using the free energy landscape in Table 7.1. The ranks of free energy differences sorted in the descending order are listed in column Rank (vaccine) and in column Rank (circulating) for the amino acids in the vaccine strain and the dominant circulating strain, respectively.

Year	Substitution	$\Delta\Delta G$ (kcal/mol)	Rank (vaccine)	Rank (circulating)
1972	T155Y	7.254 ± 0.358	14	9
1972	G158E	8.584 ± 0.479	17	6
1972	S159C	7.923 ± 0.324	17	6
1972	E190V	-9.115 ± 0.310	3	5
1973	T160A	4.160 ± 0.217	18	11
1973	N188D	19.767 ± 0.367	14	1
1973	S193N	8.239 ± 0.301	12	10
1975	S157L	-6.256 ± 0.394	15	19
1975	A160T	-4.160 ± 0.217	11	18
1975	Q189K	9.484 ± 0.640	10	2
1975	N193D	7.046 ± 0.317	10	7
1984	E156K	-26.536 ± 0.429	1	15
1984	V163A	-0.243 ± 0.217	15	16
1984	D190E	-18.752 ± 0.324	1	3
1984	I196V	-0.768 ± 0.343	16	18
1987	Y155H	-4.782 ± 0.414	9	11
1987	E188D	9.669 ± 0.382	3	1
1987	K189R	-9.872 ± 0.697	2	11
1996	V190D	27.867 ± 0.299	5	1
1996	L194I	-6.914 ± 0.324	13	17
1997	K156Q	13.140 ± 0.413	15	3
1997	E158K	-10.187 ± 0.515	6	18
1997	V190D	27.867 ± 0.299	5	1
1997	L194I	-6.914 ± 0.324	13	17
1997	V196A	5.947 ± 0.229	18	11
2003	H155T	-2.472 ± 0.355	11	14
2003	Q156H	-20.028 ± 0.365	3	20
2003	S186G	0.132 ± 0.275	14	13

from roughly 10 M to 10 M , one amino acid substitution that

epitope B.

The binding constant between hemagglutinin and antibody after the first round of maturation is about 10^6 M^{-1}, and the binding constant of an uncorrelated antibody is below 10^2 M^{-1} [72]. On average, four substitutions in epitope B change the substituted hemagglutinin sufficiently so that the immune response of the original antibody binding to epitope B is abrogated [2]. Since this is a reduction of the binding constant from roughly 10^6 M^{-1} to 10^2 M^{-1}, one amino acid substitution that contributes to immune escape causes on average a 10-fold decrease in antibody binding constant, or equivalently $\Delta\Delta G_{\text{crit}} = 1.42$ kcal/mol at 310 K. Assuming the effect of immune evasion can be broken into the sum of individual amino acid substitutions in the dominant epitope [2], we define the virus fitness w as the sum of the contribution in each site of epitope B

$$w = A_0 + \sum_{\text{epitope B}} \delta w_i. \tag{7.27}$$

We denote by $\Delta\Delta G_i^{\alpha\gamma}$ the free energy difference to substitute amino acid α to amino acid γ at site i. We investigated two versions of the virus fitness landscape. The first is to define δw_i as a linear function of the free energy difference of the substitution

$$\delta w_i = A_1 \frac{\Delta\Delta G_i^{\alpha\gamma}}{\Delta\Delta G_{\text{crit}}}. \tag{7.28}$$

The second is to define δw_i as a step function

$$\delta w_i = A_2 H\left(\Delta\Delta G_i^{\alpha\gamma} - \Delta\Delta G_{\text{crit}}\right) \tag{7.29}$$

in which H is the Heaviside step function. Illustrated in the simulation below, either definition of fitness is sufficient to explain the observed immune evasion of the H3N2 virus.

7.4.2 Selection in the Epitope

Evolution of the H3N2 virus is driven jointly by neutral evolution and selection [186]. Neutral evolution may be ongoing in sites outside the epitopes. The high substitution rate in epitope B suggests that selection is the major factor shaping the pattern of evolution in that epitope [4]. Shown in Table 7.4 and Table 7.5 are the historical substitutions. The significantly increased ranks of free energy differences suggests the existence of selection by the immune pressure for substitutions that have increased the free energy difference $\Delta\Delta G$ and decreased the antibody binding constant. The immune selection is directional: certain types of amino acids such as charged ones were initially more likely to be added into the epitope B [132] because they maximally decreased the antibody binding constant as indicated in Table 7.2. The heterogeneity of the expected free energy difference of a random substitution in Table 7.3 shows that each site in epitope B has a specific weight with regard to immune escape.

Table 7.4 also illustrates that the immune selection did not necessarily pick the amino acid with the highest rank of $\Delta\Delta G$ as the substituting amino acid. Amino acids with moderate rank were introduced into epitope B even for the fixed substitution T155Y. Therefore the historical evolution did not simply substitute amino acids by maximizing the free energy differences in Table 7.1. This phenomenon is possibly due to two causes. First, the virus fitness may be insensitive to the $\Delta\Delta G$ values, e.g. A_1 in equation 7.28 may be small, or amino acid substitutions with large $\Delta\Delta G$ values may contribute equivalently to the fitness, as in equation 7.29. Second, only a small fraction of virus in one host is shed by the host and infects the next host, so the population size of propagated virus from one host is smaller by several orders of magnitude than the total virus population size in the same host. Additionally, a seasonal bottleneck exists in the influenza virus circulation [32]. Both random mutation and

small population sizes lead to dramatic randomness in the evolution. Consequently, the evolution of H3 hemagglutinin is not solely determined by maximizing the free energy differences in Table 7.1 and minimizing the antibody binding constant, even if the virus is under immune selection. Instead, randomness plays a key role in the H3N2 virus evolution.

7.4.3 A Picture of the H3N2 Virus Evolution

Selection depends on the fitness of each virus genotype that is quantified as a non-decreasing function of the free energy difference $\Delta\Delta G$. Moderate selection in epitope B requires that fitness improvement is limited when $\Delta\Delta G$ is large. One possibility is that the ratio A_1/A_0 in equation 7.28 is small. Another is that the fitness takes the form of equation 7.29 in which all substitutions with $\Delta\Delta G > \Delta\Delta G_{\text{crit}}$ have equal fitness.

The virus evolution is also affected by the genetic drift. Genetic drift is a term which captures the random component of evolution due to the large size of the phase space of possible substitutions relative to the single set of substitutions that lead to the highest viral fitness. The effect of genetic drift is quantitatively reflected in the fixation process of a new strain, as shown in the simulation below. A narrow bottleneck of virus propagation allows only a small fraction of the progeny to survive, imposing a notable probability that a favorable substitution is lost in the next generation. The effect of genetic drift is to increase the randomness in the virus evolution so that observed substitutions are based on chance in addition to the fitness of these substitutions.

To model the H3N2 evolution discussed above, we ran two Monte Carlo simulations of the influenza evolution model. A population of N sequences of epitope B with 21

sites were created and initialized as the wildtype A/Aichi/2/1968 sequence. Here $N = 10^3$ to account for a narrow genetic bottleneck of hemagglutinin and for tractability of the simulation. We iterated the simulation program for 5,000 generations or about five years to recreate a pattern of evolution similar to that in history and shown in Table 7.4. The random substitution rate of H3 hemagglutinin is roughly 4.5×10^{-6} amino acid substitution/site/generation [37]. We let the number of substitutions follow a Poisson distribution with mean $\lambda = 21 \times 4.5 \times 10^{-6} N = 9.5 \times 10^{-5} N$ and randomly assigned the substitution sites. The substituting amino acid at each substitution site was randomly picked from the remaining 19 amino acids proportional to the historical frequencies observed in hemagglutinin. The fitness w in the first simulation was calculated for each sequence using equation 7.28 with $A_0 = 100$ and $A_1 = 3$ and that in the second simulation was calculated for each sequence using equation 7.29 with $A_0 = 100$, $A_2 = 9$, and $\Delta\Delta G_{\text{crit}} = 1.42$ kcal/mol. Note that by choosing $A_1 = 3$ for the first simulation, a random substitution causes the expected fitness to change from 100 to 104.9, and by choosing $A_2 = 9$ for the second simulation, a random substitution changes the expected fitness from 100 to 105.0. The size of the progeny of each sequences equals the fitness w of the sequence if $w > 0$, and equals 0 if $w \leq 0$. The next generation of sequences was initialized by randomly sampling N sequences from the progeny sequences.

The results of both simulations showed remarkable similarity to the observed substitutions in Table 7.4 with the bottleneck N equal to 10^3. See Figure 7.3 and Figure 7.4. Amino acid substitutions generated in the simulation are usually distinct with those in Table 7.4 observed in history. The $\Delta\Delta G$ values of each substitution emerging in the simulation are nevertheless similar to those of the historical substitutions listed in Table 7.4. As was observed in history in Table 7.4, most of the substituted strains

in the simulations with relative frequency greater than 1% have positive $\Delta\Delta G$ values with the ranks of the substituting amino acids ranging from 1 to 10. The fixation of a newly emerged substitution takes about 1,000 generations or one year on average. Fixed substitutions mostly introduce amino acids with positive $\Delta\Delta G$ values in Table 7.1 and higher ranks in Table 7.2, and several of these fixed substitutions in simulation, such as E190D and N188D, have the highest $\Delta\Delta G$ values in the current site. However, fixed substitutions in the simulation are not always the substitutions with the highest $\Delta\Delta G$ values in Table 7.1. These observations suggest that the Monte Carlo simulation considering the effect of substitution, selection, and genetic drift is able to reproduce the pattern of evolution observed in history. This simulation also shows that besides the free energy difference of each substitution, the mapping from the free energy landscape to the fitness landscape as well as the random genetic drift are dominant factors of the evolution in virus epitopes.

Shown in Figure 7.3 and Figure 7.4 for both simulations are the trajectories of relative frequencies of substituting amino acids. The trajectories are similar to historical observations of human H3N2 virus data [4]. For influenza, 1000 generations roughly equal one year. The two substitutions T155Y and N188D were fixed in epitope B in 1968–1973. As indicated by Figure 7.3 and Figure 7.4, substitution T155Y emerged between generation 3000 and 4000, or equivalently between 1971 and 1972 from the emergence of the H3N2 virus in 1968 [4]. Substitution T155Y was fixed between generation 4000 and 5000. Similarly, substitution N188D emerged between generation 2000 and 3000 and was fixed between generation 4000 and 5000. The first simulation in which virus fitness is calculated using equation 7.28 generated two fixed substitution, G129A that emerged at generation 4000 and was fixed by generation 5000, and E190D that emerged at generation 3600 and was fixed by generation 3900.

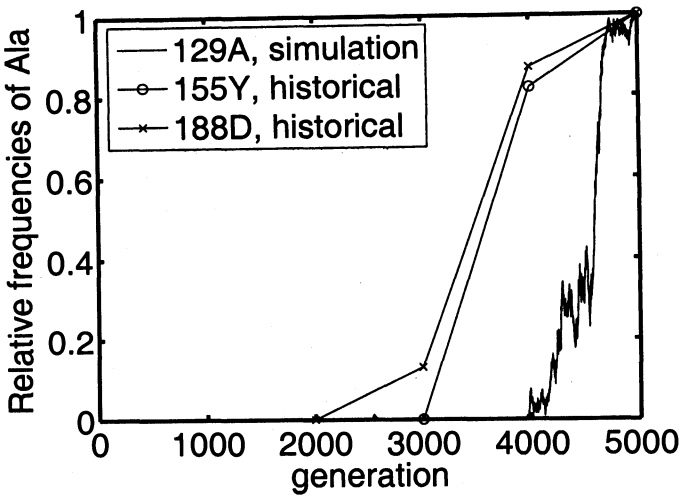
The second simulation using equation 7.29 generated one fixed substitutions, V196D emerging at generation 2900 and fixed by generation 5000, and one substitution that nearly fixed, N188D emerging at generation 4100 and acquiring the relative frequency 0.84 at generation 5000. The trajectories in both simulations resemble those of substitutions T155Y and N188D observed in history. From these results, the two Monte Carlo simulations appear to capture the main factors of immune selection and genetic drift in evolution of the H3N2 virus.

7.4.4 Multiple Substitutions

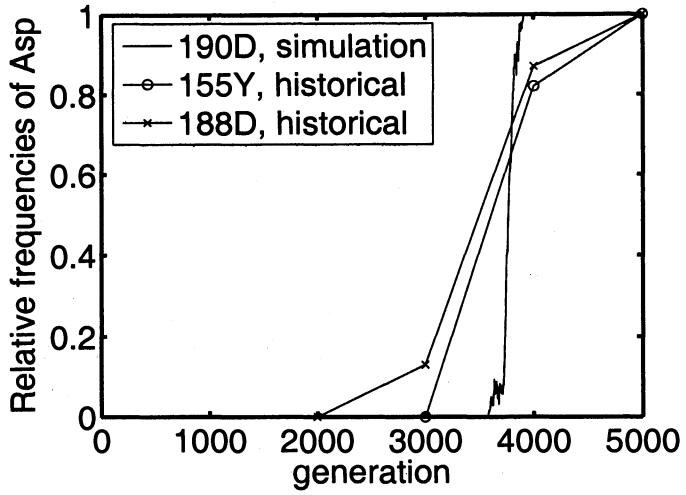
In this work, we calculated the free energy difference for each possible substitution in epitope B. The free energy calculation for multiple substitutions is intractable using the current technology due to the combinatorially increasing calculation load for multiple substitutions. The issue of multiple substitutions is here addressed by assuming that the effect of immune evasion is well represented by the sum of the contribution in each substituted site in epitope B. Data indicate the independence of the immune evasion effect of the sites in epitope B [2]. We may, thus, assume that the free energy difference of the multiple substitution is the sum of the individual $\Delta\Delta G$ values available in Table 7.1 plus a minor correction term.

7.4.5 Prediction of Future Virus Evolution

The result of this work quantifies the reduction of the binding constant of antibody to virus for substitutions in epitope B with larger $\Delta\Delta G$ values and higher ranks of substituting amino acids. A newly emerging virus strain with larger antibody binding free energy difference has a greater probability to become the dominant strain in the next flu season. Note that due to random fluctuations in the large phase space

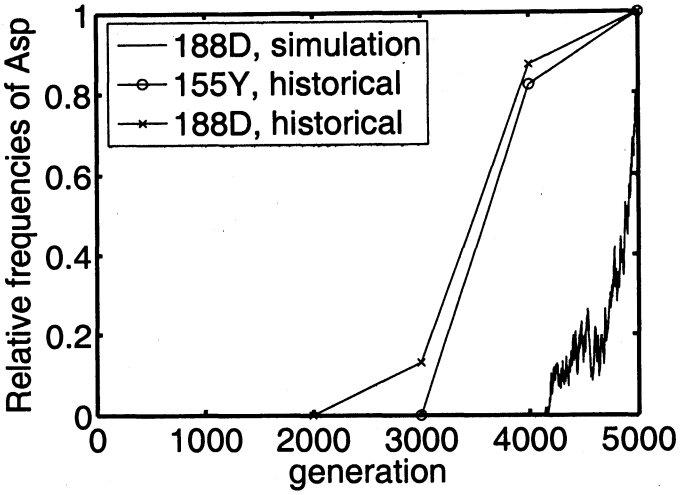


(a)

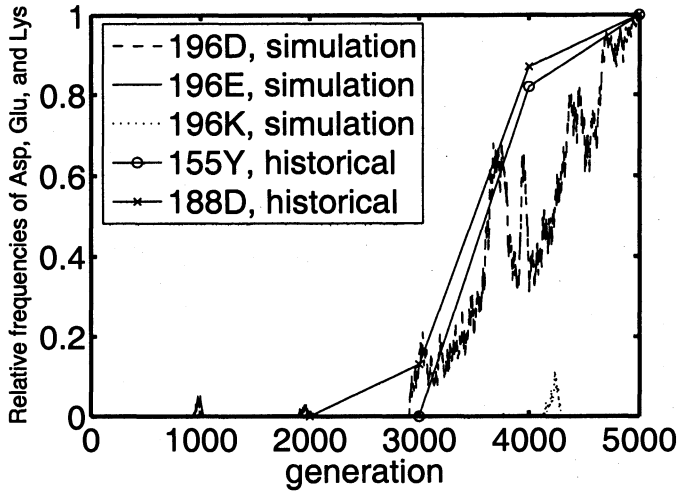


(b)

Figure 7.3 : Two fixed substitutions G129A and E190D generated by Monte Carlo simulation of epitope B using equation 7.28. Also plotted are two historical fixed substitutions in epitope B: T155Y fixed between 1971 and 1973, and N188D fixed between 1970 and 1973. The frequency data of historical substitutions are from Shih et al. [4]. The origin of time axis is 1968. One thousand generation of the H3N2 virus is approximately one year. Figure 7.3(a) Substitution G129A causing the free energy difference $\Delta\Delta G = 3.33 \pm 0.29$ kcal/mol is fixed by the simulation. The rank of the free energy difference of G129A is 12 in 19 possible substitutions in site 129. Figure 7.3(b) Substitution E190D with $\Delta\Delta G = 18.75 \pm 0.32$ kcal/mol. The rank is 1 in 19 possible substitutions in site 190.



(a)



(b)

Figure 7.4 : Two fixed substitutions N188D and V196D generated by Monte Carlo simulation of epitope B using equation 7.29. Two historical fixed substitutions T155Y and N188D are also plotted, and data are from Shih et al. [4]. Figure 7.4(a) Substitution N188D causing the free energy difference $\Delta\Delta G = 19.77 \pm 0.37$ kcal/mol is fixed by the simulation. The rank of the free energy difference of N188D is 1 in 19 possible substitutions in site 188. Figure 7.4(b) Substitution V196D with $\Delta\Delta G = 9.25 \pm 0.34$ kcal/mol. The rank is 5 in 19 possible substitutions in site 196. The proportions of substituting amino acids are represented by different line types.

of possible substitutions, actual trajectories deviate from the trajectory determined by choosing sites and substituting amino acids with greatest free energy differences. With a three dimensional structure of hemagglutinin of the current circulating virus and binding antibody, one is able to calculate the free energy landscape for all the possible single substitutions in the dominant epitope and estimate the *a priori* escape probabilities in the next season. The dominant circulating influenza strain usually possesses amino acid substitutions from the vaccine strain against which memory antibodies are generated. Usually these substitutions disrupt the antibody binding process by decreasing the binding constant, as shown in Table 7.5. Thus one can predict vaccine effectiveness by evaluating the antibody binding constant against the dominant circulating strain, which is acquired by calculating free energy difference of the amino acid substitutions between the vaccine strain and the dominant circulating strain [2]. More accurate predictions of evolutionary pattern of virus as well as epidemiological data such as vaccine effectiveness may be obtained by optimally mapping the free energy landscape to the fitness landscape and taking into account random factors such as genetic drift in the evolution process.

7.5 Conclusion

We introduced the Einstein crystal as a technology to improve the results of free energy calculation. By calculating the free energy difference of each amino acid substitution, we obtained the free energy landscape for substitutions in epitope B of hemagglutinin. There is notable variation between the values of free energy differences of different substitutions at different sites, because the identities of original and substituting amino acids, as well as the locations of amino acid substitutions, affect to differing degrees the antibody binding process. In this free energy landscape, we sug-

gest that virus tends to evolve to higher $\Delta\Delta G$ values to escape binding of antibody. Counterbalancing this selection is random drift. Historical amino acid substitutions in epitope B and Monte Carlo simulations of the virus evolution using the free energy based virus fitness, in which random genetic drift of the virus adds statistical noise into the virus evolution process, showed that selected substitutions are biased to those with positive $\Delta\Delta G$ values.

Chapter 8

Evidence for Recombination Contributing to the Evolution of Extended Spectrum β-Lactamases (ESBLs) in Clinical Isolates

In bacteria, the contribution of recombination to mosaic gene structures that facilitate the development of antibiotic resistance is well known. The main evolutionary mechanism behind the development of extended spectrum β-lactamase antibiotic resistance has long been considered to be accumulation of single point mutations, with recombination cited as a mechanistic facilitator of increased mutation rates. In this study, the presence of recombination was detected in the evolution of antibiotic resistance in clinical bacteria carrying extended spectrum β-lactamase genes by several bioinformatics methods (DNASP, LDhat, Reticulate, the Max Chi Squared test, the Sawyer's Runs Test, PhylPro, the PHI Test, and the filter method). The methods of recombination detection were calibrated against data from statistical mechanics simulations of evolution, which generated sequences under selection. Since certain strong selection pressures can mimic the effects of recombination, which the extended spectrum clinical isolates experience due to antibiotic selection at specific active sites, detection methods that rely upon silent polymorphic sites to avoid such false positive detection were also used. Additionally, linear and logistic regressions were used to further test the reliability of the results. Our findings suggest that recombination plays an evolutionary role in the development of antibiotic resistance by not only facilitating but also combining the mutations.

8.1 Introduction

The world has witnessed a sharp increase in resistant strains of bacteria [187]. Even though there are guidelines now in place for administering antibiotics to prevent over prescription, many patients, in order to reassure themselves, prefer antibiotic treatment as opposed to no treatment [187]. Resistant bacteria are now a significant problem confronting the global public health [187]. More worrying are the predictions that this problem is only to grow due to decreased research for discovery of novel and effective antibiotics [187]. We are now at a point where there are no treatments for certain kinds of infections. Factors limiting the introduction of new antibiotics include the high cost of development, significant market competition, and demographical changes within the United States [187].

Among the most frequently used antibiotics are the β-lactam drugs, which include the penicillin and cephalosporin families [188]. The most common method of resistance to this type of antibiotic is by bacteria carrying a gene that encodes for β-lactamase, thereby producing an enzyme that hydrolyzes antibiotic by cleaving its amide bond [188]. This gene can be on a plasmid, which means it can be acquired even from distant strains of bacteria through transformation, transduction, and conjugation. The two β-lactamase families TEM and SHV are among the most common in Gram-negative bacteria. We looked at variants of TEM and SHV β-lactamase genes from clinical bacterial strains. For these studies all TEM and SHV sequences in GenBank were obtained, and then only those that were observed in clinically resistant isolates (as specified by the Lahey Website) were used for the analysis. It is presumed that the original β-lactamase genes were acquired from soil bacteria [189]. The TEM and SHV families of resistance genes were both originally found in *E. coli* and *Klebsiella pneumoniae* [190]. They are now found in nearly all bacterial species

of *Enterobacteriaceae* [191].

TEM is one of the most prevalent plasmid-mediated resistance enzymes in Gram-negative bacteria (a class A β-lactamase) [188]. The TEM-1 enzyme is capable of efficient hydrolysis of penicillins and many cephalosporin antibiotics and therefore is a widespread cause of resistance. The extended-spectrum cephalosporin antibiotics such as cefotaxime and β-lactamase inhibitors such as clavulanic acid were developed to avoid the action of β-lactamases such as TEM-1. During the past 20 years, however, variants of TEM-1 β-lactamase have emerged that are able to hydrolyze extended spectrum cephalosporins or avoid the action of inhibitors [192]. There are now approximately 180 variants of the TEM-1 enzymes (termed ESBLs, or extended-spectrum β-lactamases) that provide resistance. The sequences of TEM ESBLs have been collected at the Lahey clinic ESBL website.

SHV had been considered a chromosomally encoded species-specific enzyme in Klebsiella pneumoniae [193, 194]. As another class A β-lactamase, SHV is 60% identical to TEM in the nucleotide sequence, 67% in the protein sequence, and 81% identical at the surface residues. It also has a gene that is 861 base pairs long and mutations in the gene for SHV have been discovered that lead to hydrolysis of extended spectrum cephalosporins [189, 193].

Due to the presence of β-lactamase genes in bacteria conferring varying degrees of resistance to antibiotics, it has been thought that transition from one level of resistance to another level comes about by accumulation of single point mutations [191]. More recently, it was shown that bacteria under severe stress for survival have a mechanism to increase diversity through increased mutation rate also known as the SOS response [195]. This mechanism leads to an increase in diversity, which in turn makes the population more likely to include mutations that can be selected

by the pressures of the new environment. Even though it has been established that recombination plays a role in the mechanisms that bring about the higher rates of mutation in stressed bacteria [195, 196], other (non-mechanistic) contributions of recombination have not been investigated as thoroughly. To our knowledge there has been very little recombination characterization of clinical strains carrying the β-lactamase resistant genes [197], and this research extends the work of others on recombination.

One of the requirements for recombination between TEM genes is that the genes be present in the same bacterial cell. TEM is one of the most common plasmid encoded β-lactamases in gram negative bacteria and so it is commonly found in clinical isolates [198]. The bla_{TEM} gene has been found on many different types of incompatibility group plasmids [199] and different plasmid types are often found within the same cell [200, 192]. Several studies have shown that multiple bla_{TEM} genes can be found in the same bacterial strain and different ESBL alleles can also be found [201, 202, 203]. Thus, the co-existence of bla_{TEM} alleles allows the opportunity for recombination to take place between genes in gram negative bacterial pathogens.

Our previous publications have demonstrated the power of recombination to accelerate evolution [86, 85, 204, 205]. In scenarios where the protein used by the organism is traveling a short distance in the fitness landscape to overcome the selection pressures, it is presumed that accumulation of single point mutations due to DNA replication can happen fast enough that recombination may not play a significant role in the evolutionary development. However, there are four features that suggest recombination may play a role in β-lactamase evolution.

First, β-lactamase mediated resistance has been transferred into infectious bacteria from soil strains by transformation, transduction, and conjugation [206, 207, 208].

These mechanisms all utilize recombination.

Second, specific mutations in 8 sites, generally resulting in two alleles at each site, give rise to about 30 different varieties of TEM (out of 256 theoretically possible combinations) and specific mutations in 5 sites, generally resulting in two alleles at each site, give rise to about 15 different varieties of SHV β-lactamase genes (out of 32 theoretically possible combinations). If the evolution of the β-lactamase genes were by independent random mutations and did not involve recombination, one would expect independent evolutionary trajectories. The assumption of independent, recombination-free random mutations is incompatible with the observation that a few mutations come together in many different combinations. The degeneracy of the TEM and SHV sequence space can be explained by the presence of recombination in the evolution of both proteins.

Third, recombination can bring together advantageous mutations in various combinations so that evolving into local minima can be avoided. Additionally, gradual accumulation of mutations should not create the exhaustive combination of mutations that is observed at certain sites in the genes of these clinical ESBL strains.

Fourth, certain strains of bacteria, which can carry multiple β-lactamase genes on conjugative plasmids with multiple copy numbers, exchange plasmids in the wild and allow for free recombination [207]. In strains of *Neisseria meningitidis* and *Streptococcus pneumoniae*, the rate in which alleles are altered by recombination is up to 10 times as high as the rate in which alleles are altered by point mutation [209]. As a comparison, in the same strains the rate in which individual nucleotides are altered by recombination is up to 80 times as high as the rate in which individual nucleotides are altered by point mutation [209].

The above factors suggest recombination could play an evolutionary role in ad-

dition to the mechanistic role of giving rise to mutations in stressed bacteria [194]. Strong selection on the enzyme for changes in activity leads to evolutionary characteristics that mimic recombination and makes it hard to distinguish the evolutionary paths taken. This strong selection poses the biggest challenge to establish an evolutionary role for recombination distinct from that of mutation in the development of extended-spectrum antibiotic resistance. Nevertheless, there are genetic signatures that can determine whether multiple mutations have come together via recombination or gradual accumulation. In this study we try to shed light on this issue in order to establish the evolutionary contributions of mutation and recombination.

In our previous publications, we have elaborated on the power of recombination to speed up evolution in biology [86, 85, 204, 205]. This characteristic of recombination is of immense importance because assuming infinite time, many different mechanisms can produce the desired characteristics in the population. However, since being able to evolve at a faster pace than the competition is of essence to the survival of a population, the ability to increase the pace of evolution through recombination is vitally important. Nevertheless, establishing presence of recombination in clinical isolates that have evolved in the wild without any of the controls that are available in a lab setting, using only sequence data, is a challenging task. In addition to recurrent mutation mimicking recombination signatures and creating false positive detections, lack of definitive ancestral trees for the β-lactamase genes families, and the length of the gene being only 861 nucleotides long in bacterial genes significantly confound the detection problem.

In the Results section we use different recombination detection algorithms to obtain the recombination parameter estimations, p-values, and graphical representations for the clinical data sets. In the Materials and Methods section we describe the char-

acterization of each algorithm.

8.2 Results

For this study we employed a suite of best available programs to detect recombination in the evolution of antibiotic resistance in clinical bacteria carrying the TEM and SHV extended spectrum β-lactamase genes. The TEM and SHV genes used in this study are listed in Table S1 in the supporting information. The following seven programs were employed: DNASP, LDhat, Reticulate, the Max Chi Squared algorithm, the Sawyer's Runs test, Phylpro, and the PHI test, in which the Max Chi squared algorithm and the Sawyer's Runs test are implemented in the software START; the PHI test is implemented in the software SplitsTree. Later we will describe the methodology, assumptions, and claims of the tests performed by these programs. Since certain strong selection pressures can mimic the effects of recombination [193], we made sure to include detection methods that relied upon silent polymorphic sites that are devoid of such selection pressures. We also used linear and logistic regression in addition to other methods in order to further test the reliability of our results, particularly for programs that had great sensitivity along with a higher number of false positives. The results are summarized in Table 8.1 and 8.2. We find significant recombination in both the TEM and SHV gene sets. Below we provide more details of each of the recombination detection algorithms.

For better understanding of the results, the software programs used to detect recombination were calibrated against data from statistical mechanics simulations of evolution able to generate genetic sequences with known characteristics. The algorithms were characterized using a statistical mechanics based simulation [85, 210] capable of evolving populations under various parameter values for mutation and re-

Table 8.1 : Detection of recombination in TEM and SHV

Software	DnaSP	LDhat	Max Chi Squared	Sawyer's Runs Test	PHI Test	Filter
TEM	$R = 44.4$	$R = 22$	124 recombined pairs	$p = 0.04$	$p = 0.15$	$p = 0.02$
SHV	$R = 10000$	$R = 100$	121 recombined pairs	$p = 0.03$	$p = 0.07$	N/A

The Sawyer's Runs Test is thought to be the gold standard. The table below shows either values for the recombination parameter R, the number of detected recombined pairs, or the p-value for the null hypothesis of no recombination.

Table 8.2 : Summary of simulation characterization results showing the sensitivity and the specificity of an algorithm based on our evaluation using simulated datasets

Software	DnaSP	LDhat	Reticulate	START		PhylPro	SplitsTree: The PHI Test	Filter
				Max Chi Squared	Sawyer's Runs Test			
Sensitivity	Sensitive	Sensitive	Sensitive	Sensitive	Insensitive	Insensitive	Insensitive	Insensitive
Specificity	Nonspecific	Nonspecific	Nonspecific	Nonspecific	Specific	Specific	Nonspecific	Specific
TEM Recombination	Positive	Positive	Positive	Positive	Positive	Positive	Negative	Positive
SHV Recombination	Positive	Positive	Positive	Positive	Positive	Positive	Negative	N/A

Also shown is whether the algorithm detected recombination in the clinical data. The characterization gave the sensitivity, quantified by true positives detected as a fraction of all positives in the dataset, and the specificity, quantified by true negatives detected as a fraction of all negatives in the dataset, of each algorithm.

combination (see Materials and Methods). The characterization gave the sensitivity, quantified by true positives detected as a fraction of all positives in the dataset, and the specificity, quantified by true negatives detected as a fraction of all negatives in the dataset, of each algorithm.

8.2.1 DnaSP

Introduction DnaSP, DNA sequence polymorphisms, is a population genetics based program that has algorithms allowing for several measurements within and between populations [211, 212]. The recombination test for the program is based on an algorithm that generates the rate of recombination (R), as well as the lower bound estimate for the recombination rate (Rm) [213, 214]. The program provides two recombination parameters R and Rm with values ranging between 0 and 10,000, with one being an estimated recombination rate and the other being the minimum rate of recombination that would be required to generate the population characteristics. In addition, the program provides the mutation rates (Θ), and this allows us to obtain a recombination to mutation ratio. Furthermore, the D, D', two other estimates for recombination rate (R and $R2$) are calculated under the linkage disequilibrium algorithm of the software [212]. The algorithm assumes an infinite site model, constant population size, random mating, and no population structure. The algorithm disregards sites with recurrent mutations and has no selection component.

Clinical data sets DnaSP provides estimates of various population parameters based on the sequence data. If recombination is present one should expect to obtain a non-zero rate for recombination with this algorithm and the higher the estimated recombination rate the more certain the presence of recombination. For TEM we

obtained $R = 44.4$, $Rm = 14$, and $\Theta = 8$, and for SHV we obtained $R = 10,000$ (maximum rate available by the program), $Rm = 11$, and $\Theta = 7$.

Algorithm characterization DnaSP is sensitive but nonspecific. DnaSP showed 100% sensitivity with 75 simulated data sets where recombination was present. We expected that the false positive detection for this program may be high since our simulation and clinical data did not meet the main underlying assumptions of the program, which are infinite site approximation, neutral evolution, and constant population size. To our surprise, DnaSP avoided false positive detections except when faced with high selection on specific active sites. This result would make DnaSP an excellent program for detection of recombination if one is studying datasets that have not undergone strong active site fitness selection. One can likely exclude recombination if the results from this test are not positive, although recombination can never be ruled out even if it is not detected. Unfortunately, the clinical data have been under strong active site fitness selection, hence we needed to rule out false positive detection using our regression methods and did so as explained in subsection Linear and Logistic Regression Filters.

8.2.2 LDhat

Introduction The LDhat program is based on population genetics methods [215]. It implements the approximate likelihood algorithm with some modifications [215]. It is adapted to a finite-site model, can use both haplotype as well as genotype data, estimates variable recombination rates [216], gives the option of including tri/tetraplasmic sites, and tolerates an arbitrary amount of missing data. This software also produces measurements of linkage disequilibrium.

Clinical data sets LDhat has three recombination rate estimates, the first two estimating R and Rm as described in subsection DnaSP and the third being the moments method calculation of recombination rate, as well as the Watterson's definition of mutation rate. For TEM we obtained $R = 22$, $Rm = 16$, R(moments method) $= 36$, and Θ(Watterson) $= 22$. For SHV $R = 100$, $Rm = 10$, R(moments method) $= 1738$, and Θ(Watterson) $= 16$. Once again, if recombination is present one should expect to obtain a non-zero rate for recombination with this algorithm, and the higher the estimated recombination rate the more certain the presence of recombination.

The program itself also does a randomization test of data by shuffling segments of the sequences as an additional measure of reliability by comparing various parameter values before and after randomization to examine the likelihood of the data set characteristics happening by coincidence. The randomization results showed that first, a reduction in LKmax from 178590 to 0.411; second, a reduction of G4 from 38855 to 0.3720; third, an increase in corr $(r2, d)$ from 0.01381 to 0.767; and fourth, an increase in corr (D', d) from 0.00563 to 0.4050. As it can be seen, there are very significant changes to the estimates of every parameter once the sequences are randomized, indicating that the sequence characteristics of the clinical data sets are not random. Furthermore linkage equilibrium (corr ~ 0) has been indicative of free recombination in the literature [217].

Algorithm characterization The LDhat is sensitive but nonspecific. Characterization of LDhat showed that the algorithm has good sensitivity, and has multiple outputs, including LKmax, G4, corr $(r2, d)$, and corr (D', d), which if all confirm each other give a more robust conclusion. Unfortunately LDhat also had a false positive detection weakness under strong active site fitness selection. Here again the program

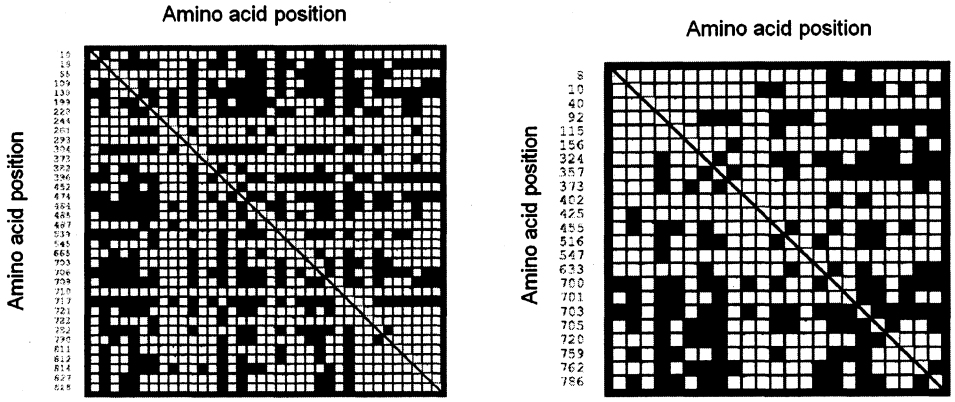
indicated the presence of recombination. The false positive needs to be detected with our regression methods.

8.2.3 Reticulate

Introduction Reticulate uses a sites patterns analysis approach [218]. Reticulate evolution refers to inter-dependent evolution among different lineages, which can come from recombination or inter-speciation, which is not relevant to the populations of our study. The program generates a matrix of pairwise compatibilities and provides a statistical analysis of the clustering obtained from compatible sites. The algorithm computes a Neighbor Similarity Score, does not have any underlying assumptions, and can do a sliding window analysis to locate recombination hot spots.

Clinical data sets For both TEM and SHV, the Reticulate algorithm displays a graphical matrix output as shown in Figure 8.1, with the black cells indicating that there are polymorphic sites that cannot be explained by maximal parsimony trees incorporating mutation without recombination. The higher the portion of incompatible cells, the more significant presence of recombination is within the population, and the higher the recombination rate. A number of incompatible cells were observed in the matrices generated by the Reticulate algorithm for both TEM and SHV, as shown in Figure 8.1. Black cells account for 35.1% and 33.6% of the non-diagonal cells in the Reticulate output in Figure 8.1 for TEM and SHV, respectively.

Algorithm characterization Characterization of the Reticulate algorithm found it to be sensitive but nonspecific. Again, we needed to rule out false positive using our regression methods and did so as explained in subsection Linear and Logistic Regression Filters.



(a) (b)

Figure 8.1 : Reticulate output for TEM (a) and SHV (b) respectively. The locations of the polymorphic sites in the genes are labeled on the axes of the squares. The black cells indicate genetic patterns that cannot be explained by point mutation as the sole mechanism of evolution.

8.2.4 The Max Chi Squared Algorithm and the Sawyer's Runs Test

Introduction Pairwise sequence comparisons are the bases of both the Max Chi Squared test and the Sawyer's Runs test [219, 220, 221]. The Max Chi Squared test is statistical analysis of clustering of polymorphic sites for a pair of sequences. The Sawyer's Runs Test is based on gene conversion analysis in a set of sequences explained in [221]. This test relies on silent mutation sites for its analysis. It looks to determine if there are more consecutive identical silent polymorphic sites in different regions than that predicted by general probabilities. Both of these programs do not have any underlying assumptions; furthermore, the Sawyer's Runs Test is based on silent mutation sites that are entirely unaffected by selection pressures.

Clinical data sets for the Max Chi Squared algorithm The Max Chi Squared algorithm produces p-values for recombination between a pair of sequences by shuffling different segments of the sequences many times and examining how many of the random shuffles show equal or greater recombination signatures than the original pair. For TEM we observed that 124 pairs of sequences had recombination present with p-values below 0.05, and we observed 121 pair with p-values below 0.05 for SHV.

Algorithm characterization for the Max Chi Squared algorithm The characterization of the Max Chi Squared algorithm found it to be sensitive but nonspecific. The Max Chi Squared algorithm detected recombination with p-values less than 0.05 for all the 75 simulated data sets with recombination. Nevertheless, it allowed for false positive detection both under strong active site fitness selection as well as general selection. Again the program indicates presence of recombination. The false positive with strong active site fitness selection needed addressing using the filter (general

selection was ruled out by previous software programs).

Clinical data sets for the Sawyer's Runs test With the Sawyer's Runs test we obtained p-values of 0.04 and 0.03 for recombination presence in TEM and SHV respectively. The detection of recombination with statistically significant p-values by this method is important for three reasons: First the recombination signal was strong enough to overcome the insensitivity of the program (see simulation characterization below). Second the method utilizes silent mutations largely devoid of selection pressure. Third if the domain of allowable mutations is extended beyond the silent polymorphic sites to all polymorphic sites (including non-silent ones) the p-values go above 0.88 for TEM and 0.60 for SHV. The last factor shows that if one looks for recombination based on all polymorphic sites, one could miss the recombination signature in the evolution of antibiotic resistance, due to heavy selection pressures destroying the recombination signatures.

Algorithm characterization for the Sawyer's Runs test The characterization of the Sawyer's Runs test showed it to be very insensitive but specific. The characterization gave p-values around 0.5 for simulation generated datasets that employed recombination. Consequently, we did not obtain any false positive detection using this algorithm. The Sawyer's Runs test has 100% specificity with 25 simulated data sets where recombination was absent.

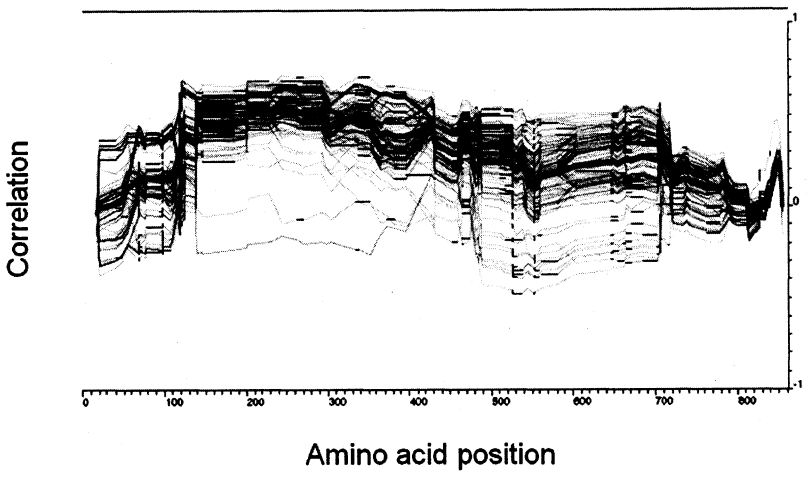
8.2.5 PhylPro

Introduction PhylPro, Phylogenic Profile Algorithm, is a graphical pairwise sequence comparison program [5]. It uses a sliding window algorithm to analyze the agreement of patterns of distances in particular regions. Even though graphical rep-

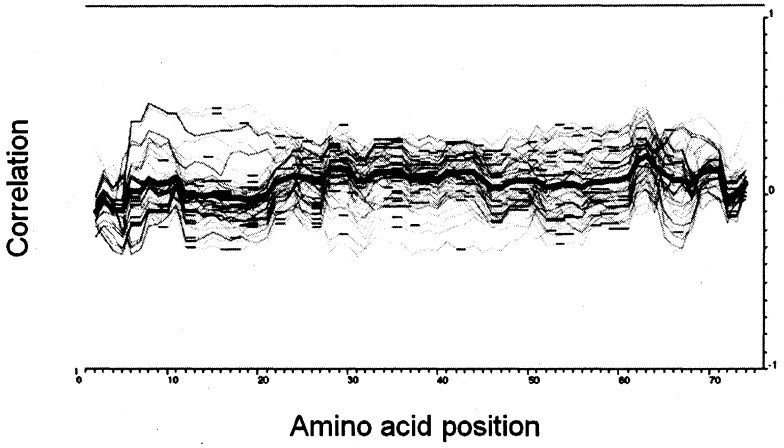
resentations are able to give cues for patterns that are hard to quantify, this lack of quantification can lead to difficulty of interpretation of the results. Nevertheless, the graph is based on pairwise correlations and recombination destroys the correlation signal. Hence the evidence for recombination becomes more definitive as the square or absolute value of results nears zero. The documentation does not mention any underlying assumptions for its algorithm.

Clinical data sets PhylPro produces a graphical pairwise sequence comparison that displays the agreement (correlation) of patterns of distances using a sliding window analysis. It was the analysis [217] that guided us to use near zero correlation of pairwise comparisons, which happened to be the case for TEM and SHV (shown in Figure 8.2), as indicating free (or very high) recombination in the TEM and SHV genes [217].

Algorithm characterization The PhylPro program was insensitive and specific. The pattern generated by PhylPro indicated existence of recombination in all the 50 simulated data sets with recombination present. As for the PhylPro algorithm, characterization showed some distinguishing characteristics between the false positives and the true positives, but due to the non-quantitative nature of the program it was hard to identify a cut off value for distinguishing between true and false positives. The PhylPro program does not provide a quantitative measure of the recombination, such as the p-value in the Max Chi Squared algorithm and in the Sawyer's Runs Test. We therefore had to identify recombination by inspecting the figure generated by the PhylPro.



(a)



(b)

Figure 8.2 : PhylPro results for TEM (a) and SHV (b) respectively. Each black curve represents a phylogenetic profile generated by PhylPro [5]. The red curve is the mean of all the phylogenetic profiles. In independent evolutionary trajectories that would come about in a mutation-only evolution process, the mutations in each particular trajectory are strongly correlated to each other. Hence the graph for the correlation measure for a gene pool that has evolved under a mutation-only scenario would be away from the Correlation = 0 line. Recombination destroys this correlation between different polymorphic sites, causing the graph to significantly deviate from the Correlation = 1 line. The higher the recombination rate, the further the plot deviates from the Correlation = 1 line and moves towards the Correlation = 0 line. (a) PhylPro results for TEM are close to the Correlation = 0 line at many positions. These low correlations can come from high recombination rates or low selection pressure [6]. The selection of TEM is usually intensive in the presence of antibiotics. Therefore, these low correlations suggests high recombination rates of TEM. (b) Similarly, PhylPro results for SHV are close to the Correlation = 0 line at all the positions, suggesting even higher recombination rates of SHV.

8.2.6 The PHI Test

Introduction The PHI test, which is the acronym of the Pairwise Homoplasy Index test [222], is performed by the software SplitsTree [223], a phylogenic network analysis software [224, 225]. The algorithm is a site pattern analysis using incompatibility scoring that provides p-values for recombination presence. This is a general test with no underlying assumptions, which claims to discern between recurrent mutation and recombination.

Clinical data sets With the PHI test algorithm we obtained p-values of 0.15 and 0.07 for recombination presence in TEM and SHV respectively. The p-values being above 0.05 was attributed to the insensitivity of the program, as discussed below. We regarded the p-values obtained as positive indications of recombination since they were two of the lowest p-values that we observed with the PHI test, indicating that the recombination in the clinical datasets were far stronger than those generated through our simulations.

Algorithm characterization The PHI test algorithm was characterized to be insensitive and nonspecific. The PHI test yielded p-values generally around 0.50 even with recombination present. It was considered as very positive indication that some of the lowest p-values obtained by us using the Sawyer's Runs test and PHI test were those of the TEM and SHV. However, 3 out of 25 simulated data sets where recombination was absent showed p-values less than 0.05. The PHI test is hence nonspecific.

8.2.7 Linear and Logistic Regression Filters

Linear and logistic regression filters were employed, using the R software package for statistical analysis, in an attempt to classify the false positives and true positives in different categories and use these findings for the clinical data sets. In doing so we tried using the values of the following factors as our input of parameters: tetraplasmy (existence of four different/degenerate alleles at the same nucleotide site), triplasmy (existence of three different/denerate alleles at the same nucleotide site), polymorphic sites, non-informative polymorphic sites, informative polymorphic sites, mutations, tetraplasmy divided by triplasmy, tetraplasmy divided by polymorphic sites, tetraplasmy divided by non-informative sites, tetraplasmy divided by informative sites. The reason behind our original selection of factors related to tetraplasmy was that under high selection the mutant enzymes get selected and fixed within the population almost immediately. After mutations get fixed, the population goes under quasi-neutral evolution. Since recombination increases diversity in neutral evolution [226, 227] we suspected that tetraplasmy should be higher in populations where recombination plays a role. The regression resulted in mean values for simulation sequence sets with recombination (true positives) and simulation sequence sets without recombination (false positives) that were statistically different (p-value of 1.17×10^{-27}). The classification rejects 98% of false positives while keeping 20% of true positives or alternatively rejects 80% of false positives while keeping 40% of true positives. TEM was classified as true positive under both versions of the test (SHV only carried triplasmic variants; hence this test was not applicable). The logistic and weighted logistic regressions produced similar results without significant differences.

8.3 Discussion

The theoretical importance of recombination and horizontal gene transfer to the emergence of modularity under changing environments and to increased adaptation rate in biology has been emphasized [86, 85, 204, 205]. The emergence of modularity is reflected by the increased proportion of intramodule connections in the three dimensional protein structure [86]. Evidence for the contribution of recombination to mosaic gene structures has been presented [228]. Our findings in this chapter address the situation where the divergent genes cannot be attributed to any one family of bacteria, and a chronological family tree cannot be established among ESBL mutants.

All but one software program detected recombination (Table 8.2). Only the extremely insensitive PHI test did not detect recombination in TEM and SHV. The results were further reinforced with linear and logistic regression to make sure the results are not due to false positive detection. Considering that these sequences are picked by evolutionary selection and obtained from clinical settings, any signature from recombination detected in these datasets shows that recombination has already played a direct role in the evolution of antibiotic resistance development. Since we started this study, the p-values of the Max Chi Squared Test, of the Sawyer's Runs Test, and of the PHI Test have gone down steadily with the addition of new sequences in GenBank. This gives us extra confidence that with additional data those p-values that are higher than the 0.05 threshold may fall below this threshold in time as more data points are added. Another measure that we took to reinforce our findings was to remove a progressively larger region of the genes that included the active sites, and observed the effect on the recombination signature. Under these conditions, we still detected recombination for both the TEM and SHV families, though the magnitudes of the recombination coefficients were smaller to varying degrees in the TEM, SHV,

and simulation datasets.

Evolution by point mutation alone leaves some results unexplained. First, recombination is known to take place in plasmids on which the resistance genes are located. Second, in one case there were three strains with a single point mutation difference from the original TEM-1 strain each [193, 229], and there also existed all the eight possible combinations of these three mutations in the population in an exhaustive manner. The evidence for recombination was so strong that the authors of [229] proposed combination, without using the term recombination, as a strong possibility for obtaining the combinatorial variants. We examined the population in order to see how many strains could be obtained by simply combining other strains with single mutations together. Combining five strains with a single mutation difference from TEM-1 each (TEM-2, TEM-12, TEM-17, TEM-19, and TEM-29) gives rise to 11 new strains that are already in the population (TEM-3, TEM-6, TEM7, TEM-11, TEM-15, TEM-16, TEM-18, TEM-26, TEM-44, TEM-129, and TEM-134) providing us with a diversity of 17 from the exhaustive 32 possible variants of these 5 mutations in clinical isolates. This high combinatorial diversity suggests the existence of recombination in the evolution from the original TEM-1.

One of the major implications of our results is the potential for increase in diversity of β-lactamase resistance in short timeframes, already observed in clinical settings. Though this complements our prior publications [86, 85, 204, 205], as well as those from other groups [226], it is not agreed upon by all [230]. An increase in the population diversity leads to a higher propensity for evolution of resistance in bacteria carrying these genes. This implies that despite different mutation combinations having varying resistance levels, the overall population can contain most, if not all, of them to acquire future adaptability. This is of particular importance because it

allows the bacteria to adapt to local fitness loci without being trapped there and still be adaptable to future environmental/antibiotic changes. Recombination can explain results that otherwise seem unlikely [231]. A calculation of the 99% probability density of evolutionary pathways for TEM-1 to become resistant to cefotaxime predicted 10 pathways [231], yet some of the clinically observed data were not present in those pathways. This result suggests the presence of recombination. Recombination can explain these inconsistencies by allowing for diversity beyond the fittest strains and allowing for combining of the mutations in many different ways. With recombination, it is important to notice that the bacteria seem to increase their resistance to certain antibiotics to moderate levels without being trapped in local minima that can reduce their adaptability in the future.

Since we used β-lactamase sequences from clinically resistant bacteria that have been evolutionarily selected in real life settings, our findings complement the results from [226], as well as those from [197]. Their previous studies had shown for the β-lactamase genes in particular that with random mutation the functionality of the new strains declines exponentially with accumulation of mutations due to lack of conservation of the protein fold. Recombination in β-lactamase genes, however, not only avoids an exponential decline, but also has a symmetrical and log-parabolic curve, giving rise to "broad sequence diversity with relatively low cost in loss of function" [226]. In the present work, we showed that organisms take advantage of such benefits by using recombination for their evolution even over short mutational distances in the fitness landscape. In our previous publications [86, 85, 204, 205] we had shown that the greatest contribution from recombination is increasing the evolution rate of a population. Our previous findings along with the new observations in clinical datasets and the present results lead us to conclude that recombination, due to its acceleration

of evolution in populations, conservative nature in preserving function under the right conditions, creation of broad diversity, and exchange of genetic information among different organisms, is a significant contributor to evolution. Recombination brings mutations together, contributes mechanistically to generate mutations, and creates gene rearrangements and amplification.

8.4 Materials and Methods

The statistical mechanics based simulation software was developed by our group [85, 210]. The program was modified to correspond to our simulation needs in this project. We set the sequence length to 900 nucleotides. Mutations were performed at the nucleotide level, with fitness a function only of amino acid sequence. Therefore, the protein structure contains 300 amino acid residues. We set the ratio of recombination to mutation to range from 0 to 10, with a quarter the results at a ratio of 0 and half performed at 1, and the last quarter at 10. We defined the active site, the part recognizing the substrate for ESBL, by randomly choosing P amino acids with P the size of the active site. We set the strength of selection based on active site, which measures the degree of selection on the active site residues, to have a weight of 1 to 50, with a ratio of 5 chosen for final simulations. Population sizes of 1000 were used. These parameters mimic the strong selection pressure observed with the clinically observed TEM and SHV enzyme variants [210]. The simulations went through 30,000 rounds of replication and selection to generate the data sets. The simulations were used to characterize the recombination detection software programs. We used a generalized NK model to simulate the fitness of the protein enzyme. The active site fitness constant of enzyme and target is of the following form: $\exp(a - b\langle U\rangle)$, where a and b are constants, and we define U, a measure of fitness of interaction where the

lower the value of U is the fitter the enzyme has become, for the enzyme by:

$$U = \sum_{i=1}^{M} U_{\alpha_i}^{\mathrm{sd}} + \sum_{i>j=1}^{M} U_{ij}^{\mathrm{sd-sd}} + \sum_{i=1}^{P} U_i^{\mathrm{c}} \tag{8.1}$$

$$U_{\alpha_i}^{\mathrm{sd}} = \frac{1}{\sqrt{M\left(N-K+1\right)}} \sum_{j=1}^{N-K+1} \sigma_{\alpha_i}\left(a_j, a_{j+1}, \ldots, a_{j+K-1}\right) \tag{8.2}$$

$$U_{ij}^{\mathrm{sd-sd}} = \sqrt{\frac{2}{DM\left(M-1\right)}} \sum_{k=1}^{D} \sigma_{ij}^{k}\left(a_{l_1}^{(i)}, \ldots, a_{l_{k/2}}^{(i)}; a_{l_{k/2+1}}^{(i)}, \ldots, a_{l_k}^{(i)}\right) \tag{8.3}$$

$$U_i^{\mathrm{c}} = \sigma_i\left(a_i\right)/\sqrt{P} \tag{8.4}$$

Here the number of enzyme secondary structures is $M = 30$ (For our case to be comparable to the clinical sequence length we chose $M = 30$, making the length $M \times N = 300$.) The number of enzyme amino acids in the active site is $P = 5$. The number of amino acids in a single secondary structure sub-domain is $N = 10$, and range of local interactions is $K = 4$ amino acids. There are $L = 5$ different sub-domains in a secondary structure (helices, strands, loops, turns and others) represented by the subscript α_i. The number of interactions between secondary structures is $D = 6$. The first summation is over all the interactions within a secondary structure. The second term represents the inter-secondary structure interactions. The total number of interaction that a typical amino acid can have is roughly 12, and $[2\left(k-1\right)]$ of them are within secondary structure sub-domain and $[D\left(M-1\right)/N]$ are in the inter-secondary structure sub-domain interactions. For sequence length of $M \times N = 300$, the simulation is a semi-quantitative model of a protein domain. The third summation represents the active site measure of fitness and its fitness is represented by a reduction in this parameter. The sites considered for the active site fitness parameter represented the active site residues in the TEM and SHV structures. This allows for simulation study of the evolution of active site residues under different conditions. Also there is a strong correlation between the active site fitness and

enzyme activity in clinical data [232].

Initially we generated 5 optimized high fitness sub-domain pools, each made of 300 variants corresponding to $L = 5$ types to have biologically relevant fitness levels for the structures to choose from. Out of each set of 300 we pick 3, then going on to generate 1000 random sequences by different combinations of these segments for 100 different secondary structure positions in each sequence. From this initial population, the top 50% is selected based on our fitness formulation. The next step is to repopulate back to the original count, and allow for mutation and recombination, with parameters set for any specific scenario, among the new generation. The selection and repopulation representing one generation cycle is repeated for 30,000 generations. Further details are in [210]. These sequences are used to characterize the software programs that were chosen to analyze our clinical sequences.

The simulations are appropriate for such characterizations due to following reasons: the statistical mechanics GNK model used in the simulation is uncorrelated to the detection algorithms that are used in the software programs, the semi-quantitative nature of the simulations makes the sequences mimic protein structures closely, and the mutation and recombination parameters are adjustable for verification of the recombination detection ability of the software programs under different conditions.

We used the simulation algorithm described above to generate sequences by point mutation, recombination, and selection to characterize the recombination detection algorithms, and better understand and interpret the results obtained from them. We generated 25 sequences with the ratio of recombination to mutation set to 0, 50 sequences with the ratio of recombination to mutation set to 1, and 25 sequences with the ratio of recombination to mutation set to 10. To detect recombination, we used the following seven programs: DNASP, LDhat, Reticulate, the Max Chi Squared

algorithm, the Sawyer's Runs test, Phylpro, and the PHI test. Note that the Max Chi Squared test and the Sawyer's Runs test are performed by the software START (Sequence Type Analysis and Recombination Tests) [219]. Additionally, the PHI test is performed by the software SplitsTree [223].

Chapter 9

A Two-Scale Model for Correlation in B Cell VDJ Usage of Zebrafish

The zebrafish (*Danio rerio*) is one of the model animals for study of immunology because the dynamics of the adaptive immune system in zebrafish are similar to that in higher animals. In this work, we built a two-scale model to simulate the dynamics of B cells in primary and secondary immune reactions in zebrafish and to explain the reported correlation between VDJ usage of B cell repertoires in individual zebrafish. We use a delay ordinary differential equation (ODE) system to model the immune responses in the 6-month lifespan of zebrafish. We use the generalized NK model to simulate the B cell maturation process in the primary or the secondary immune response to a single type of antigen within 10 days. The generalized NK model shows that mature B cells specific to one antigen largely possess a single VDJ recombination. In addition, the probability that mature B cells in two zebrafish have the same VDJ recombination increases with the B cell population size or the B cell selection intensity and decreases with the B cell hypermutation rate. The two-scale model shows a distribution of correlation in the VDJ usage of the B cell repertoires in two six-month-old zebrafish that is highly similar to that from experiment. This work shows that the spin glass theory of the immune response, in combination with a long-time ODE model, accurately describes aspects of the immune system dynamics in zebrafish.

9.1 Introduction

B cell-mediated adaptive immunity exists in jawed animals [76]. B cells protect hosts by secreting antibodies that recognize and neutralize pathogens and foreign substances. Immunity generated by B cells is hence indispensable to the hosts' survival. The primary immune response occurs when a novel type of antigen is detected by the immune system. The antigen is processed and presented to naïve B cells, which mature in the germinal center. In the maturation process, the B cells acquire the capability to recognize and neutralize a specific antigen. First, a naïve B cell recombines one V gene segment, one D gene segment, and one J gene segment in the genome to create the nucleotide sequence encoding the antibody. Second, this nucleotide sequence undergoes multiple rounds of somatic hypermutation, and B cells with high affinity to the antigen are selected. The selected mature B cells differentiate into the antibody-secreting plasma cells or long-lived memory B cells that effectively activate the secondary immune response against the same antigen in the future. Janeway *et al.* provide a more detailed review of the B cell-mediated immune reaction [72]. Understanding the dynamics of and relationship between VDJ recombination and somatic hypermutation informs one about the central mechanism of B cell immunity.

Recent experimental studies provide information on the B cell maturation process in zebrafish. Zebrafish (*Danio rerio*) have been increasingly used as a model animal to study the immune system because experiments on zebrafish are easy to perform, zebrafish reproduce quickly, and zebrafish possess one of the most primitive adaptive immune systems, which is a model for the adaptive immune systems in humans and mice [77, 78, 79]. The genome of zebrafish contains 39 V gene segments, five D gene segments, and five J gene segments, which together encode the V region of immunoglobulin IgM heavy chain in zebrafish [80, 81]. High-throughput sequencing

of the complete IgM repertoires in 14 six-month-old zebrafish revealed that one fish carries up to 5000–6000 distinct nucleotide sequence of IgM with $39 \times 5 \times 5 = 975$ VDJ recombinations [7]. VDJ usage, the probabilities that each of the 975 possible VDJ recombinations is used in the IgM repertoire, has a correlation coefficient between individual zebrafish up to $r = 0.75$ [7].

In the present study, we developed a two-scale model to illustrate that B cell maturation processes in distinct individuals, even though random, may converge to the same VDJ recombination in the same environment of antigens. We use delay ordinary differential equations (ODEs) to simulate the immune response against multiple antigens circulating in the environment. We use the generalized NK model to describe B cell maturation processes against one antigen. The original NK model builds a random rugged energy landscape on which peptides evolve [82, 83]. The parameters of the NK model for short peptides have been fit to the observed data. Mora *et al.* fit a random energy model, similar to the NK model, with a large number, approximately 10^3, of parameters to experimentally measured probabilities of D gene segment usage in zebrafish [84]. As an extension of the original NK model, the generalized NK model takes into account the interaction between distinct subdomains of a protein and protein-protein interaction [85]. The generalized NK model can describe maturation of the whole V region of antibodies [34] and evolution of proteins in general [86]. In this study, we use the generalized NK model to analyze the convergence in the VDJ usage of immune response of two fish exposed to the same set of antigens. We extracted results of the generalized NK model to assign VDJ recombination to each type of mature B cells, the dynamics of which are solved by the ODE model. Correlation coefficients between the VDJ usage calculated from theory agree with experiment, in which most pairs of zebrafish had a weak correlation with $r \leq 0.2$ and

a small fraction of pairs had $r > 0.5$ [7].

The two-scale model is motivated by the nature of immune response. The adaptive immune system receives different antigens at various time points. An antigen can initiate a primary or secondary immune response, depending on whether the infected individual has seen the antigen before. During the immune response B cells undergo rounds of somatic hypermutation and selection. The ODE system models at a mean field level the dynamics of B cell repertoires [87, 88]. The generalized NK model computes the repertoire of B cells that respond and evolve in response to the antigen. The generalized NK model explicitly simulates the somatic hypermutation and selection of the B cell repertoire reacting to a specific type of antigen in one individual [85, 34]. We combined the ODE model and the generalized NK model to build the present two-scale model, a model that can zoom out to yield a global view of the dynamics of B cell repertoires and zoom in to focus on the somatic evolution of the B cells reacting to one type of antigen.

The purpose of this study is to model the B cell-mediated immune response and to explain the correlation in the VDJ usage in distinct individuals. This model is able to describe the mechanism of B cell-mediated immunity and to depict a snapshot of B cell repertoires at any time point in the host's life span. The Materials and Methods section illustrates the ODE model for multiple types of antigen and the generalized NK model for a single type of antigen. The Results and Discussion section shows the simulation results at both scales and compares the simulation results to experimental data. Finally, we present our conclusions and outlook.

9.2 Materials and methods

The model simulates the B cell-mediated immune response in zebrafish living in an environment with multiple types of antigen. This model was built in two scales. At the first scale, the model describes the B cells involved in immune response against multiple types of antigen. A delay ODE model simulates the maturation process of multiple clones of B cells against distinct antigens. At the second scale, the model describes B cells reacting to a single type of antigen. A generalized NK model describes the rugged energy landscape of the antibody V region in the maturation process against one antigen, the process of VDJ recombination and subsequent somatic hypermutation. Figure 9.1 illustrates the model at both scales.

The two scales of the model are connected by assigning VDJ recombinations to the mature B cells generated by two zebrafish exposed to one type of antigen, according to the probability computed by the generalized NK model. The delay ODE system computes the dynamics of the number of mature B cells in each immune reaction. The generalized NK model computes the immune responses in two zebrafish against an antigen, denoted antigen i. Due to selection, the VDJ usage in a single fish after the primary response is almost always localized to a single VDJ recombination. The generalized NK model computes the probability that this VDJ recombination is the same in both fish, $p = 0.327$, and distinct with probability $1 - p$. For each primary or secondary immune reaction induced by antigen i modeled by the ODE system, we assign the antigen i specific mature B cells one VDJ recombination for two zebrafish with probability p, and two distinct VDJ recombinations for two zebrafish with probability $1 - p$. In this way, the VDJ repertoire for two fish are constructed as a function of the antigenic environment.

9.2.1 ODE model

The delay ordinary differential equation (ODE) system computes the dynamics of the immune response triggered by antigens. The ODE system is a mean field model of the number of B cells responding to antigen over time. The number of mature B cells in a zebrafish can reach the order of magnitude of 10^3 [7]. Zebrafish lived in an environment with N_{ag} types of antigens. For one zebrafish, inoculation of antigen i triggers the immune response that boosts the number of relatively short-lived B cells and long-lived memory B cells in a zebrafish. The immune response of each zebrafish against antigen i is modeled by two delay ODEs:

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = c_1 v_i(t-\tau_1) + c_3 v_i(t)\, y_i(t) - b x_i(t) \qquad i = 1, 2, \ldots, N_{\text{ag}} \tag{9.1}$$

$$\frac{\mathrm{d}y_i(t)}{\mathrm{d}t} = c_2 v_i(t-\tau_2) \qquad i = 1, 2, \ldots, N_{\text{ag}}. \tag{9.2}$$

in which state variables x_i and y_i are the numbers of antigen i specific plasma cells, which secret antibodies, and memory B cells, respectively. The initial conditions at the time of hatching are $x_i(0) = 0$, $y_i(0) = 0$ because of lack of antigenic experience. The level of antigen i received by the zebrafish, v_i, is subject to a random process. Antigen inoculation is random, but the same for both zebrafish, since the environment is common for both zebrafish. We assume that $N_{\text{ag}} = 10$ distinct types of antigen exist in the environment. The zebrafish are in one of two states, the normal state in which the antigen is absent, and the infected state in which the antigen is present in the zebrafish. A newborn zebrafish is not inoculated by antigens and so it is in the normal state, $v_i(0) = 0$. In an antigen inoculation, both zebrafish receive one randomly selected type of antigen, denoted antigen i. The average time span between two events of infection is $\lambda = 30$ days. The value of $v_i(t)$ jumps from 0 to 100 at antigen inoculation, and the zebrafish transit from the normal state to the infected

state. The antigen presentation lasts for approximate one week [233]. Therefore, the zebrafish transit from the infected state to the normal state and v_i falls back to zero after seven days if the host was not inoculated by antigen i in the past seven days. The zebrafish stay in the infected state if they were inoculated by antigen i or another type of antigen, antigen j, in the past seven days. The ODE system described above contains $2N_{\mathrm{ag}}$ equations for both zebrafish.

When the host received antigen i at time t, a primary immune response is triggered if the host is naïve to this antigen ($y_i(t) = 0$). Otherwise a secondary immune response is triggered. Antigen i stimulates production of antigen i specific plasma cells with rate $c_1 v_i(t - \tau_1)$ and memory B cells with rate $c_2 v_i(t - \tau_2)$. In the primary immune response, B cells appear around $\tau_1 = 5$ days after the initial contact of antigen, and memory B cells appear around $\tau_2 = 30$ days after the contact [72]. In the secondary immune response, existing memory B cells mount an immediate reaction to the corresponding antigen with a rate $c_3 v_i(t)\, y_i(t)$, which is higher than the rate $c_1 v_i(t - \tau_1)$ in the primary immune response because $y_i(t) \gg 1$. The first order term bx_i in equation 9.1 quantifies the decay of antigen i specific plasma cells with rate $b = 5\ \mathrm{day}^{-1}$, because most B cells in germinal centers live a short time before apoptosis [72].

This ODE system was solved by numerical integration. The solution gives the composition of mature B cells in a zebrafish challenged by multiple types of antigen during its life history. The mature B cells at $t = 180$ days plus the naïve repertoire represents the B cell repertoire of a 6-month-old zebrafish.

9.2.2 Generalized NK model

The generalized NK model defines the rugged random potential energy landscape, in which the antibody V regions of B cells mutate randomly and are under selection [85, 34, 204, 234, 86, 235]. The generalized NK model assigns an energy value to each antibody V region with a specific amino acid sequence and a tertiary structure. The energy negatively correlates to the fitness of the corresponding B cell in the maturation process. The energy of an antibody V region is expressed by

$$U = \sum_{i=1}^{M} U_{\alpha_i}^{\mathrm{sd}} + \sum_{i>j=1}^{M} U_{ij}^{\mathrm{sd-sd}} + \sum_{i=1}^{P} U_i^{\mathrm{c}} \tag{9.3}$$

in which M is the number of secondary structural subdomains, P is the number of amino acids contributing to the antibody-antigen binding process. The energy U is the sum of three components: secondary structural subdomain energies ($U_{\alpha_i}^{\mathrm{sd}}$), subdomain-subdomain interaction energies ($\sum_{i>j=1}^{M} U_{ij}^{\mathrm{sd-sd}}$), and chemical binding energies ($\sum_{i=1}^{P} U_i^{\mathrm{c}}$). The parameters of the generalized NK model are fixed in our previous papers [85, 34].

The secondary structural subdomain energy is expressed by

$$U_{\alpha_i}^{\mathrm{sd}} = \sqrt{\frac{1}{M\,(N-K+1)}} \sum_{j=1}^{N-K+1} \sigma_{\alpha_i}\left(a_j, a_{j+1}, \ldots, a_{j+K-1}\right) \tag{9.4}$$

in which N is the number of amino acids in a subdomain and each amino acid in the subdomain interact with $K-1$ other amino acids in the same subdomain. Here $N = 10$ and $K = 4$. An antibody V region contains approximately 120 amino acids [72] or equivalently $M = 12$ secondary structures. The identity of the subdomain is denoted by α_i. There are $L = 5$ types of subdomains, which are helices, strands, loops, turns, and others, Therefore $\alpha_i = 1, \ldots, 5$. The identity of amino acid in position j is represented by a_j. The 20 amino acids are categorized into $Q = 5$ classes, which

are negative, positive, polar, hydrophobic, and other. So $a_j = 1, \ldots, 5$. For each of L types of subdomains, σ_{α_i} is a K-dimensional Gaussian array with zero mean and unit standard deviation.

The subdomain-subdomain interaction energy is expressed by

$$U_{ij}^{\text{sd-sd}} = \sqrt{\frac{2}{DM(M-1)}} \sum_{k=1}^{D} \sigma_{ij}^{k} \left(a_{j_1}^{i}, \ldots, a_{j_{K/2}}^{i}; a_{j_{K/2+1}}^{j}, \ldots, a_{j_K}^{j} \right) \tag{9.5}$$

Between subdomains i and subdomain j $(i < j)$, $D = 6$ interactions occurs. Each interaction involves $K/2$ interacting amino acids in subdomain i and $K/2$ interacting amino acids in subdomain j. For each pair of subdomains (i, j), σ_{ij}^{k} is a K-dimensional Gaussian array with zero mean and unit standard deviation.

The chemical binding energy is expressed by

$$U_i^{\text{c}} = \sqrt{\frac{1}{P}} \sigma_i (a_i) \tag{9.6}$$

in which a_i is the identity of one of the $P = 5$ amino acids in the antibody V region contributing to the binding to the antigen. The quantity $\sigma_i (a_i)$ is a Gaussian with zero mean and unit standard deviation.

Two generalized NK models with identical parameters were implemented to calculate the immune response against the same antigen in two zebrafish. In each zebrafish, the number of B cells was set to $N_{\text{size}} = 2000$ because the number of B cells in a germinal center is in the order of magnitude of 10^3 [236, 237]. The V region contains approximately 120 amino acids in which 100, 10, and 10 amino acids are encoded by the V, D, and J gene segment, respectively. Thus the generalized NK model fixes the length of V region to 120 amino acids and defines amino acid 1–100, 101–110, and 111–120 as encoded by the V, D, and J gene segment, respectively. There exist five types of secondary structures, which are helices, strands, loops, turns, and others, and each secondary structure contains around 10 amino acids [85]. For each type of

the secondary structure, we built a pool of 300 amino acid sequences with length 10 and minimized the energy of each of these sequences using Metropolis Monte Carlo method [85]. By randomly selecting and recombining 10-residue sequences from the five pools, we created 39 V segments consisting of 10 secondary structures, five D segments consisting of one secondary structure, and five J segments consisting of one secondary structure. These V, D, and J segments were randomly recombined to create the initial $N_{\text{size}} = 2000$ structures. In each round of simulation, the structures underwent somatic hypermutation and selection. The rate of somatic hypermutation is about $n_{\text{mut}} = 0.5$ amino acid/structure/generation [72]; thus we set the number of mutated amino acid in each generation in each structure to follow a Poisson distribution with mean $\lambda = 0.5$. In each generation, $p_{\text{cut}} = 20\%$ of the structures with the lowest energy were propagated to the next generation [34]. The duration of primary immune response is around 10 days, or equivalently 30 generations of B cells [233, 72]; thus we set the simulation to comprise 30 rounds of somatic hypermutation and selection. The secondary immune response comprises another 30 rounds of somatic hypermutation. We recorded the VDJ usage in each zebrafish in each round of simulation.

9.2.3 Model characterization and verification

The ODE system models the number of antigen i specific B cells in each zebrafish in a mean-field approach. The duration of primary and secondary immune response, which are explicitly modeled by the number of iterations in the generalized NK model, are respectively treated with the duration τ_1 and τ_2 in the ODE system. The generalized NK model calculates the rapid increase of antibody affinity of mature B cells in the primary immune response [34]. The generalized NK model also calculates the further

increase of antibody affinity in the secondary immune response [34]. Note that in the generalized NK model, the secondary immune response does not necessarily occur immediately after the primary immune response.

The results from this model of the immune system were compared to experiment [7]. In particular, the correlation of VDJ usage from the simulation was compared with that from the experiment. The high-throughput sequencing experiment by Weinstein *et al.* presented the IgM heavy chain frequencies of 975 VDJ recombination in each zebrafish, and obtained 91 correlation coefficients between the VDJ usage in 14 zebrafish [7]. Both the simulated and the experimentally observed correlation coefficients fall into four bins: independence (correlation coefficient $r \leq 0.1$), low correlation ($0.1 < r \leq 0.2$), moderate correlation ($0.2 < r \leq 0.5$), and high correlation ($r > 0.5$), following the categorization scheme of Weinstein *et al.* [7]. The histogram of the correlation coefficients generated by our model was compared with that from the experiment using Pearson's χ^2 test.

9.3 Results and discussion

9.3.1 Overview of the results

As described in the Materials and Methods section, the model in this study considers both the time course of the number of B cells as well as the sequence identity of those B cells, modeled respectively by the ODE and generalized NK models. The following text describes the results at both scales of the model, illustrated in figure 9.1. The generalized NK model focuses on the B cell maturation process in response to challenge by a specific type of antigen. Subsection 9.3.2 presents the results of the generalized NK model simulating the maturation process of naïve B cells against

the one type of antigen. The trajectories of the simulation give the VDJ usage as a function of time in the antibody repertoire of each zebrafish and the correlation of the VDJ usage between two zebrafish. Note that the immune system may produce mature B cells with a distinct VDJ usage against an identical antigen, because the immune maturation process is random. Subsection 9.3.3 presents the results of the sensitivity analysis of three parameters, which are the number of naïve B cells reacting to the antigen, the somatic hypermutation rate, and the selection intensity. In subsection 9.3.4, the distribution of energy changes ΔU of single mutations in the generalized NK model is presented.

The ODE model defined in equations 9.1 and 9.2 uses the antigens recognized at distinct times of the lifespan of zebrafish as the input and obtains the number of each type of B cells at one given time. The ODE model does not take into account the maturation process of somatic hypermutation. The ODE model takes into account selection in a mean field way. In subsection 9.3.5, the algorithm is used to model the challenge of antigen v_i as a Poisson process. Equations 9.1 and 9.2 are then numerically solved to acquire the dynamics of the number of each type of plasma cell and memory B cell. Finally, by combining the correlation data of each type of B cell in two zebrafish obtained by the generalized NK model and the number of each type of B cells generated by the ODE model, we present the correlation coefficient between the VDJ usage of the whole B cell repertoire in two zebrafish. Pearson's χ^2 test shows that the distribution of correlation coefficients between VDJ usage in the B cell repertoires in two zebrafish in the model agrees with those measured experimentally [7] (p-value $= 0.62$).

9.3.2 Primary and secondary immune response against one antigen

The generalized NK model simulated the dynamics of the sequence diversity of the B cells in the primary and secondary immune responses. Mutation and selection causes the B cells to be localized in both the amino acid sequence space and the VDJ usage. With the same parameters, 1000 independent runs of the generalized NK model yielded 1000 individual trajectories, which were also processed to give one average trajectory. The number of distinct sequences and the number of VDJ usage decreased as the simulation proceeded, as shown in figure 9.2(a). In $N_{\text{size}} = 2000$ naïve B cells, the average number of genotypes and VDJ recombinations were 1997.9 and 849.6, respectively. In the primary immune response, the average number of genotypes rapidly decreased from the initial value of 1997.9 to 58.8 in the first six generations and then slowly decreased to 30.2 at generation 30 in an exponential way. In the secondary immune response, the average number of genotypes decreased from the value of 30.2 to 11.4 at generation 60 following the same exponential function of time. As shown in figure 9.2(a), dynamics of the numbers of genotypes and VDJ recombinations show that localization in the VDJ recombination space occurred prior to that in the sequence space. The number of VDJ recombinations decreased at a higher speed than did the number of genotypes. The average number of VDJ recombinations reduced from the initial value of 849.6 to 1.94 in 15 generations, then decreased to 1.13 at generation 30 at the end of the primary immune response, and further to 1.003 at generation 60 at the end of the secondary immune response. Thus, the B cells converged to one dominant VDJ recombination in nearly all immune responses. A dominant VDJ recombination exists in the 2000 B cells at the end of both the primary and the secondary immune response. We calculated the fraction of B cells with the dominant VDJ recombination in each of the 1000 runs. Out of

the 1000 runs, 950 runs have the fraction greater than 0.85 at the end of the primary immune response, and over 950 runs have the fraction equal to 1 at the end of the secondary immune response. The dominant VDJ recombinations in two different zebrafish is identical at the end of the primary and the secondary immune responses with probability $p_1 = 0.326$ and $p_2 = 0.327$, respectively. Because p_1 and p_2 are similar, we defined $p = 0.327$ as the probability that two different zebrafish have the same dominant VDJ recombination at the end of an antigen-specific immune response, whether primary or secondary.

The energy of a VDJ recombination in the initial $N_{\text{size}} = 2000$ naïve B cells has a relationship to the probability that mature B cells converge to this VDJ recombination. For each of the 1000 runs, we obtained the identity of the VDJ recombination in mature B cells at the end of the secondary immune response and calculated its rank in the initial energy in the probability distribution of 975 VDJ recombinations in naïve B cells. The 1000 values of ranks fall into 10 bins that are 1–100, 101–200, ..., 900–975, as presented in figure 9.2(b). Figure 9.2(b) shows that VDJ recombinations with higher ranks in the initial energy have larger probability to be the only VDJ recombination in the mature B cells. Comparison between the observed ranks and those in a null model shows the significance of this tendency for the VDJ recombinations with higher ranks to be eventually fixed, with the p-value shown below. For each run, the null model randomly chose the VDJ recombination in one naïve B cell and calculated its rank. The 1000 randomly chosen VDJ recombinations in the null model had significantly lower ranks than those in the mature B cells (p-value $< 6 \times 10^{-24}$, Wilcoxon signed rank test). The relationship also indicates that the initial energy of VDJ recombination is correlated with the identity of the VDJ recombination that was eventually fixed in the mature B cells. As a result, the mechanism of VDJ re-

combination in two zebrafish leads to the correlation of VDJ usage in their B cell repertoires, as suggested in the previous work on zebrafish B cell repertoires [7].

The generalized NK model allows us to calculate the correlation coefficients between VDJ usage in two zebrafish of B cells against the same antigen. In each of the 1000 runs in the simulation, the correlation coefficient values in naïve B cells and B cells in each generation in the primary and the secondary immune responses show a unique dynamics. As described in figure 9.3, naïve B cells at generation 0 in two individuals show uncorrelated VDJ usage. The correlation coefficient increases rapidly to around 0.5 in the first three generations and after that showed large variation. At the end of the primary immune reaction with 30 generations, most mature B cells are evolved from the same VDJ recombination. The generalized NK model presents the correlation coefficient r between the VDJ recombinations of the two zebrafish. The probability for $r > 0.995$ at the end of the primary and secondary immune responses is 0.301 and 0.327, respectively. The probability for $r < 0.1$ at the end of the primary and secondary immune responses is 0.641 and 0.672, respectively.

The dynamics of correlation coefficients shown in figure 9.3 reflect the mechanism of B cell maturation. Limited by the local point mutation moves of somatic hypermutation, B cells only explore a small neighborhood of the VDJ recombination in the sequence space in the first three generations. Therefore the energy of B cells in the first three generations largely depends on the energy of the VDJ recombination. Selection pressure removes VDJ recombinations with high energy in the zebrafish and increases the correlation coefficient between two zebrafish. After the third generation, B cells explore broader areas in the energy landscape, and the energies of B cells depend less on the initial VDJ recombination. Thus the B cell maturation process becomes more random after the third generation. The trajectories of corre-

lation coefficients in figure 9.3 converge to $-1/975$ or 1 because the mature B cells against one antigen in most individuals show only one VDJ recombination, see figure 9.2(a). The correlation coefficient between VDJ usage in two zebrafish is $-1/975$ if the mature B cells have distinct VDJ recombinations in the two zebrafish. The correlation coefficient is 1 if the VDJ recombination is identical in the two zebrafish. Correlations $-1/975$ and 1 are, therefore, two absorbing states in the random process of correlation coefficients. Figure 9.3 then shows that the VDJ usage in B cells is localized in the primary immune response; the secondary immune responses do not change the VDJ usage in most cases.

9.3.3 Parameter sensitivity in the generalized NK model

The dynamics of B cell maturation processes simulated in the generalized NK model depend on several factors, including N_{size} as the number of B cells reacting to the antigen, n_{mut} as the average number of point mutations in each generation, and p_{cut} as the proportion of B cells propagated to the next generation. With the correlation coefficients r between VDJ usage calculated in the generalized NK model, the relative frequency that $r > 0.995$ after the secondary immune response defines the probability, p, that mature B cells in two individuals converge to the same VDJ recombination. Note that $r > 0.995$ implies almost always the same VDJ recombination in two fish. Similarly, the probability, q, that mature B cells converge to two distinct VDJ recombinations is defined as the relative frequency with $r < 0.1$ after the secondary immune response. The result of the generalized NK model shows that the first probability $p = 0.327$ and the second is $q = 0.672$. Note that the sum of the probabilities could be less than one if greater than one VDJ recombination exists in either fish.

As shown in section 9.3.2, B cells in two zebrafish from the primary immune

response against the same antigen converged to one identical VDJ recombination with probability p and converged to two distinct VDJ recombinations with probability q. There existed a small probability $1-p-q$ that VDJ recombination was not converged in either zebrafish. Figure 9.4(a) plots p, q, and $1-p-q$ as the function of N_{size}. The value of p increased rapidly from 0.207 when $N_{\text{size}} = 1000$ to 0.475 when $N_{\text{size}} = 10000$ and were insensitive to N_{size} with $N_{\text{size}} > 10000$. The probability $1 - p - q$ for non-converged VDJ recombinations increased with N_{size}.

The increasing trends of p and $1 - p - q$ come from the delayed localization of the VDJ recombination with larger N_{size}. As illustrated by figures 9.4(b) and 9.4(c), primary immune responses with greater N_{size} localize B cells in the sequence space and in the VDJ recombination space more slowly than do those with lower N_{size}. The B cells thus explore the energy landscape associated with the sequence space, along with the different VDJ recombinations, for a longer period of time. A larger value of N_{size} also enables the B cells to generate more mutants and to explore a broader subregion of the energy landscape. The B cells in two zebrafish, both of which explore the energy landscape more intensively, have a higher chance to evolve to the same local minimum in the energy landscape.

When the value of N_{size} is fixed, the values of n_{mut} and p_{cut} affect the B cell maturation process. The effect of varying n_{mut} and p_{cut} is reflected in two indices of the process, which are p defined above and N_{max} defined as the maximum number of B cells with identical sequence at the end of the primary immune response. If N_{max} is close to N_{size}, most of the mature B cells from the primary immune response share the same genotype. The values used in the generalized NK model are $n_{\text{mut}} = 0.5$ and $p_{\text{cut}} = 0.2$. Here N_{size} possesses a fixed value and both n_{mut} and p_{cut} independently vary from 0.1 to 1.0, as shown in figures 9.5 and 9.6. The values of both p and N_{max}

decrease quickly with p_{cut}. The values of p and N_{max} fall to near zero with $p_{\text{cut}} > 0.9$. Both p and N_{max} are less sensitive to n_{mut}.

The two parameters n_{mut} and p_{cut} govern the process of B cell somatic hypermutation. The values of mutation rate n_{mut} and selection pressure p_{cut} observed in experiment have evolved during the evolutionary history of the immune systems [238]. The low sensitivity to n_{mut} in figures 9.5 and 9.6 indicates that with a hypermutation rate n_{mut} on the order of magnitude 10^{-1}/sequence/generation, the immune system maintains the capability to explore the sequence space and locate the local minima in the energy landscape. On the other hand, the selection pressure p_{cut} controls the B cell exploration in the rugged energy landscape. A small value of p_{cut} confines the B cells in compact regions in the sequence space by only allowing B cell genotypes with lowest energy to propagate to the next generation. Smaller values of p_{cut} cause both zebrafish to have a higher chance to reach the same local energy minima, as described in figure 9.6, and result in increased correlation in the VDJ usage, as illustrated in figure 9.5.

9.3.4 Distribution of the energy change ΔU of a point mutation

The generalized NK model used in this work was used to compute the statistics of potential energy U and ΔU, which is the energy difference caused by a point mutation, with experimental data on the point mutation of proteins. The generalized NK model for the primary and secondary immune responses simulated the maturation of 1000 antigen-specific B cells with a 60-generation timespan. In each generation from 0 to 60, ΔU was calculated by temporarily introducing a random mutation in each B cell. The simulation was repeated 20 times. Thus, a total of $61 \times 1000 \times 20 = 1.22 \times 10^6$ values of ΔU and the corresponding energy U before the mutation were calculated.

Figure 9.7(a) describes the distribution of ΔU. The point mutations with $\Delta U < 0$ comprise 3.8% of the 1.22×10^6 mutations simulated in the generalized NK model. This fraction is 4.9% in the experimental data of point mutations affecting protein affinity [238]. Figure 9.7(b) plots the data points $(U, \Delta U)$ generated in the simulation in a two-dimensional space. The linear regression between U and ΔU produces a trend line:

$$\Delta U = -0.027U + 0.109. \tag{9.7}$$

The slope of this trend line is significantly different from zero (p-value $< 2.2 \times 10^{-16}$). The negative slope is expected because ΔU is expected to have a symmetric distribution with center zero when U equals to zero in the generalized NK model [85, 34]. Equation 9.7 is consistent with this expectation within error bars. However, figure 9.7(b) also illustrates that the correlation between U and ΔU is weak ($R^2 = 0.018$). This weakly correlated U and ΔU were also observed in the experimental data of affinity-related amino acid mutations in the PINT database [238].

9.3.5 Co-effect of multiple types of antigens

The generalized NK model analyzes the immune response against a single type of antigen. In reality, the zebrafish are challenged by various types of antigens in their lifespan. Each type of antigen induces a distinct immune response producing specific mature B cells. First, the numerical solution of the delay ODE system defined by equations 9.1 and 9.2 gives the dynamics, which derive the B cell repertoire at time t, of each type of mature B cells. Second, the results from the generalized NK model assign the identity of VDJ recombinations to each type of antigen-specific mature B cells in each zebrafish. The correlation coefficients of VDJ usage in the B cell repertoires in two zebrafish agree with those observed in experiment [7].

The solution of the ODE model defined by the equations 9.1 and 9.2 is determined by the presence of each antigen i denoted by $v_i(t)$, $t = 0$–180 days. In the simulation, we first randomly selected antigens to which the zebrafish are exposed at several time points. The dynamic system defined by equations 9.1 and 9.2 was numerically solved using the parameters $c_1 = 1$, $c_2 = 0.1$, and $c_3 = 0.03$. The physical meanings of these parameters are given in section 9.2.1. Figure 9.8 shows the dynamics of the total plasma cells and the total memory B cells, respectively, determined by one run of the simulation. Each peak of plasma B cells corresponds to an antigen inoculation. For each instance of the simulation there are distinct dynamics of plasma cells and memory B cells due to the different antigenic environments. The simulation was repeated 6000 times. Each run gave the dynamics of the ODE model. The results for the 6000 runs are used to analyze the correlation of the VDJ recombinations in both zebrafish.

The correlation between the two zebrafish is illustrated by the snapshot of the composition of plasma cells and memory B cells at one time point. A previous experiment measured the VDJ usage in 6-month-old zebrafish [7] and so the snapshot time was day 180. Corresponding to the 10 types of antigen, each of the 6000 runs of the simulation generated the numbers of the 10 kinds of antigen-specific mature B cells, each of which consisted of plasma cells and memory B cells. As calculated by the generalized NK model, each type of mature B cells shares the same VDJ recombination in two zebrafish with probability $p = 0.327$. Using this result, we assigned each type of mature B cells in the two zebrafish identical VDJ recombination with probability $p = 0.327$ and distinct VDJ recombination otherwise. Besides the mature B cells, a substantial number of naïve B cells circulate in zebrafish. The number of naïve B cells in one zebrafish was fixed to 10^5, following the literature [7]. In the model, each naïve

B cell possesses a random VDJ recombination and there is no correlation between VDJ usage in the naïve B cells of two zebrafish. For each zebrafish, the model yielded the VDJ usage in each zebrafish by counting the frequency of each VDJ recombination in naïve and mature B cells. The model also calculated the correlation coefficient between the VDJ usage in two zebrafish for each run of the simulation. In total, the model presented 6000 model-calculated correlation coefficients.

The distribution of the 91 correlation coefficients between the VDJ usage in B cell repertoires in distinct zebrafish in experiment [7] and that of the 6000 correlation coefficients generated in the model are similar. We adopted the experimental categorization scheme for the correlation coefficient [7], which classifies the correlation coefficients r in four bins: no correlation with $r < 0.1$, low correlation with $0.1 \leq r < 0.2$, moderate correlation with $0.2 \leq r < 0.5$, and high correlation with $r \leq 0.5$. Figure 9.9(a) shows the relative frequency of correlation coefficients in experiment in each of the four bins. Figure 9.9(b) shows the distribution of correlation coefficients from the simulation. A high level of similarity exists between these two frequency distributions (p-value $= 0.62$, Pearson χ^2 test).

9.4 Conclusion and outlook

In this work, a two-scale model illustrates the zebrafish immune response in the presence of multiple types of antigens. Establishing the model in two scales allows incorporation of different mathematical tools with specific focuses into the system.

The first scale of the model is the set of delay ODEs, which is a mean-field approximation on the genotype of B cells specific to one type of antigen. The dynamical system defined by the ODEs receives the challenge of distinct types of antigen as the input signal. The solution of the ODE system gives the composition of the mature

B cell repertoire at any time point between 0 and 180 days. The correlation between VDJ usage in the B cell repertoires in two zebrafish is determined by the dynamics of B cell compositions and the probability p, calculated by the generalized NK model, that the same antigen induces B cells with identical VDJ usage. The correlation data of VDJ usage calculated in the two-scale model agree with those from the experiments [7].

The second scale of the model is the generalized NK model focusing on the somatic evolution of B cells specific to a single type of antigen. Each type of antigen in the environment leads to distinct memory B cell evolution. We here use the generalized NK model for each type of antigen to build the energy landscape. This generalized NK model accurately describes the somatic evolution of B cells against each antigen. The generalized NK model describes the affinity maturation process. The random walk B cells with lower energy in the rugged energy landscape have higher probability to survive and produce progeny. Selection over B cells removes genotypes and VDJ usages with high energy and hence substantially localizes the B cells in the rugged energy landscape and in the $39 \times 5 \times 5$ VDJ recombination space. VDJ usage is localized and fixed in the primary immune response. The localization, however, is not a deterministic process. Primary immune response has the probability p to yield mature B cells with an identical VDJ usage in two zebrafish, as calculated by the generalized NK model. The probability p increases with N_{size}, the number of B cells reacting to a specific type of antigen, while decreasing with n_{mut} and p_{cut}, which are the somatic hypermutation rate and the survival rate in the selection, respectively. With a large number of B cells reacting to a specific type of antigen, N_{size} B cells explore a broader region of the energy landscape and have a higher chance to find the same local minimum in the primary immune reaction in two zebrafish. A high

level of n_{mut} increases the uncertainty in the B cell maturation process. A high level of p_{cut} increases the tendency to fix the B cell in the current location in the energy landscape. Therefore the probability p decreases with both n_{mut} and p_{cut}.

The correlation between VDJ usage in two zebrafish indicates that evolution is not a completely random process. In the generalized NK model, the VDJ usage in naïve B cells is independently initialized in different zebrafish. The probability p that two zebrafish possess mature B cells with identical VDJ usage is calculated by the generalized NK model. The generalized NK model asserts that the somatic evolution is not an unbiased random walk; the selection pressure and the energy landscape affected by the specific antigen strongly drive the somatic evolution of B cells in distinct zebrafish in one identical direction. In the ODE model, correlation of the VDJ usage in the B cell repertoire on day 180 also shows that the B cell somatic evolution in two zebrafish may be correlated if they are challenged by the same antigen at the same time. Consequently, closely related antigen challenges could induce immune responses with similar characters in distinct individuals.

This two-scale model is flexible and extensible to analyze the immune system dynamics in zebrafish and higher species. The study on the zebrafish B cell repertoire can be extended by analyzing the immune system in zebrafish at different ages. The B cell repertoire in these individuals may reveal the dynamics of development of immune systems challenged by various types of antigen, and further calibrate the ODE model. Like the environments with identical challenging antigens discussed above, closely related genomes may also cause correlated VDJ usage. We propose sequencing of the B cell repertoire of zebrafish in different families living in different controllable environments to test the contribution to VDJ usage of two factors: zebrafish genome and antigens in the environment. If the correlation of VDJ usage in distinct zebrafish

strongly depends on their genetic similarity, their genomes are a determinant of VDJ usage. If not, VDJ usage is independent of the genome of each individual. The two-scale model analyzing the adaptive immune system in zebrafish, a model animal in immunity, could also be extended to study the adaptive immune response in higher species.

This model describes the mechanism of B cell maturation and humoral immunity in a quantitative way. In the maturation process, the B cells are first localized to one VDJ recombination and then further increase the binding affinity to the antigen by hypermutation and selection. VDJ recombinations with high initial affinity to the antigen prior to the somatic hypermutation have large chances to be selected in the maturation process and to be present in the mature B cells. The hypermutation and selection of the B cells are not deterministic, and so two zebrafish inoculated by the same type of antigen generate mature B cells with identical VDJ usages with probability p. The probability p increases with N_{size}, the number of B cells in the germinal center, decreases with n_{mut}, the hypermutation rate, and decreases with p_{cut}, the fraction of B cells surviving each round of selection. As the B cell maturation proceeds, available sequences with higher affinity to the antigen become rarer. This trend explains the rapid increase of affinity in the primary immune response and the relatively slower increase of affinity in the secondary immune response. The experimental data show that the theoretical description presented here matches aspects of the zebrafish immune system evolutionary dynamics.

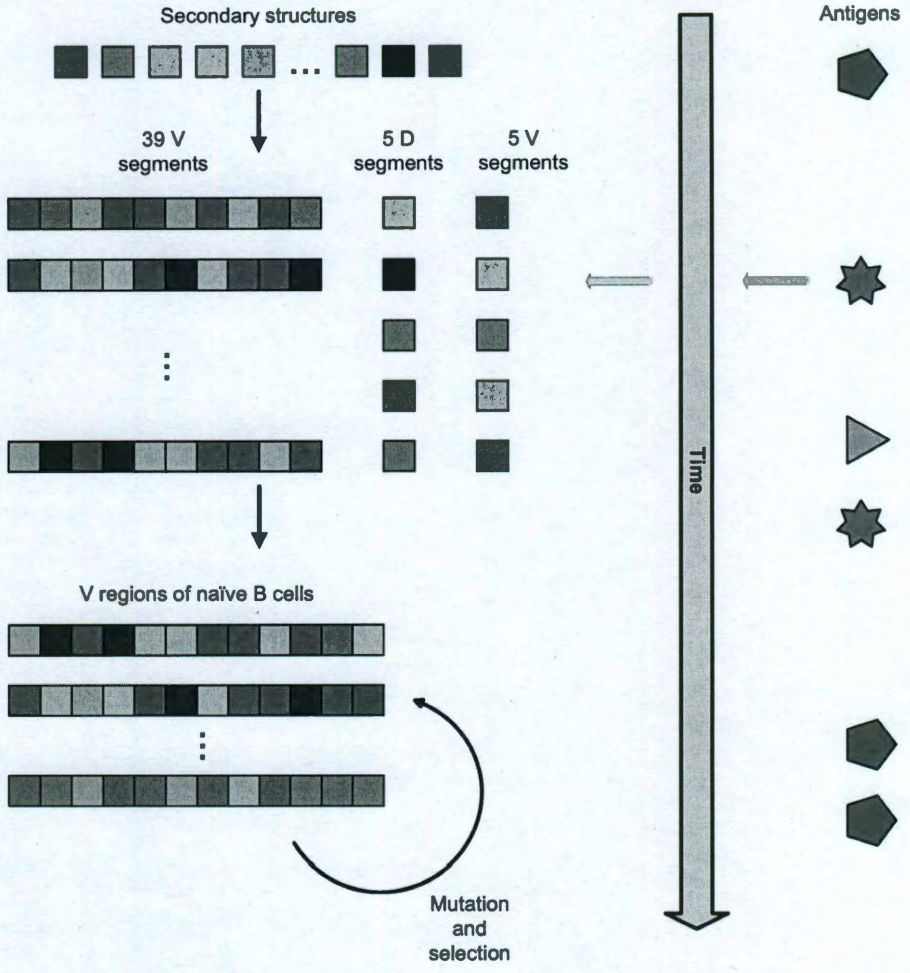
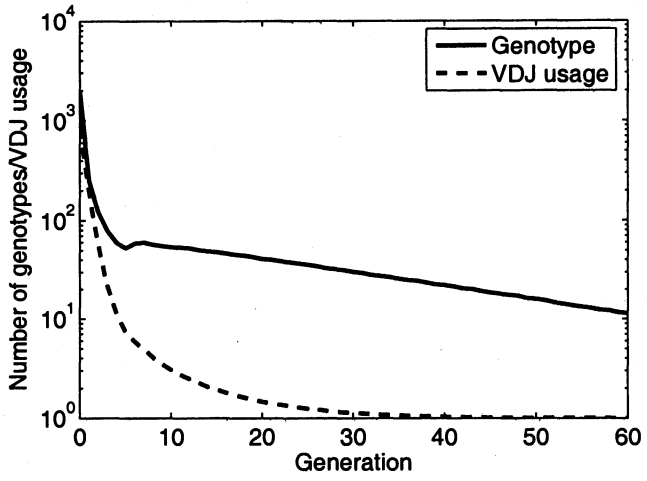
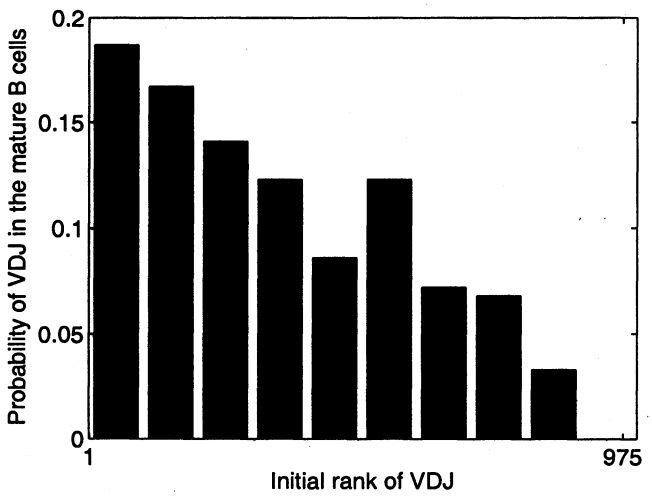


Figure 9.1 : Illustration of the two-scale model of zebrafish immune response. The timeline arrow represents the zebrafish life history from 0 (hatch) to 180 days. To the right of the timeline is the first scale. Antigen challenged the zebrafish at several random timepoints. Each challenge led to a primary or secondary immune response. To the left of the timeline arrow is the second scale of the model. The flow chart describes the generalized NK model. Distinct secondary structures represented by squares with different colors were first built by minimizing the energy using Metropolis Monte Carlo method. These secondary structures randomly recombined to form V, D, and J segments, which randomly recombined to form the V region of the IgM heavy chain. The V region underwent 30 rounds of mutation and selection in the primary immune response and another 30 rounds in the secondary immune response. In this figure, as an example, the generalized NK model describes the primary immune response against the second antigen this zebrafish met in its lifetime, which is represented by the blue star.



(a)



(b)

Figure 9.2 : (a) The numbers of distinct genotypes and VDJ recombination in the primary (generation 1–30) and secondary (generation 31–60) immune responses against one antigen involving $N_{\text{size}} = 2000$ naïve B cells. The number of VDJ recombination decreased much faster than that of B cell genotypes. In most cases all the B cells showed 1–2 VDJ recombinations at the end of the primary immune response and one VDJ recombination at the end of the secondary immune response. (b) Probability distribution at generation 60 of the rank of probabilities of $39 \times 5 \times 5 = 975$ VDJ recombination in $N_{\text{size}} = 2000$ naïve B cells reacting to one type of antigen. This probability distribution used 1000 rank data generated by running the generalized model 1000 times. The naïve VDJ ranks fell into 10 bins with rank 1–100, 101–200, ..., 901–975. Bin 10 with rank 901–975 was empty.

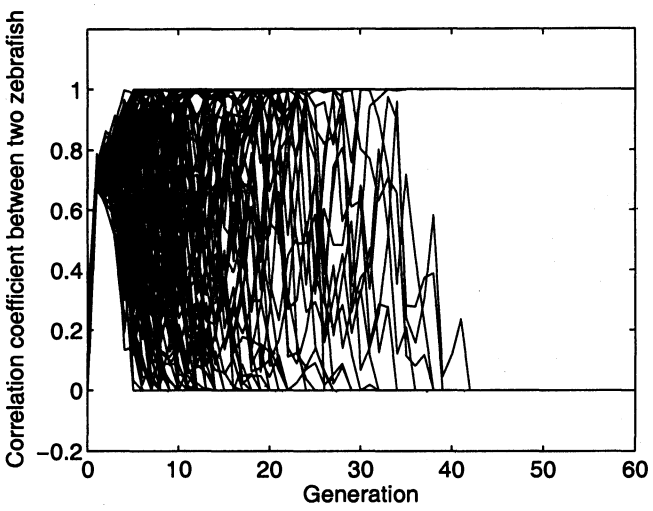


Figure 9.3 : Trajectories of correlation coefficients r between VDJ usage of B cells in two zebrafish reacting to one certain antigen. The simulation consisted of 1000 runs, each of which generated the correlation coefficient r between naïve B cells in generation 0 and their progenies in generation 1–30 in the primary immune response and in generation 31–60 in the secondary immune response. The first 100 out of 1000 trajectories are here plotted for clarity.

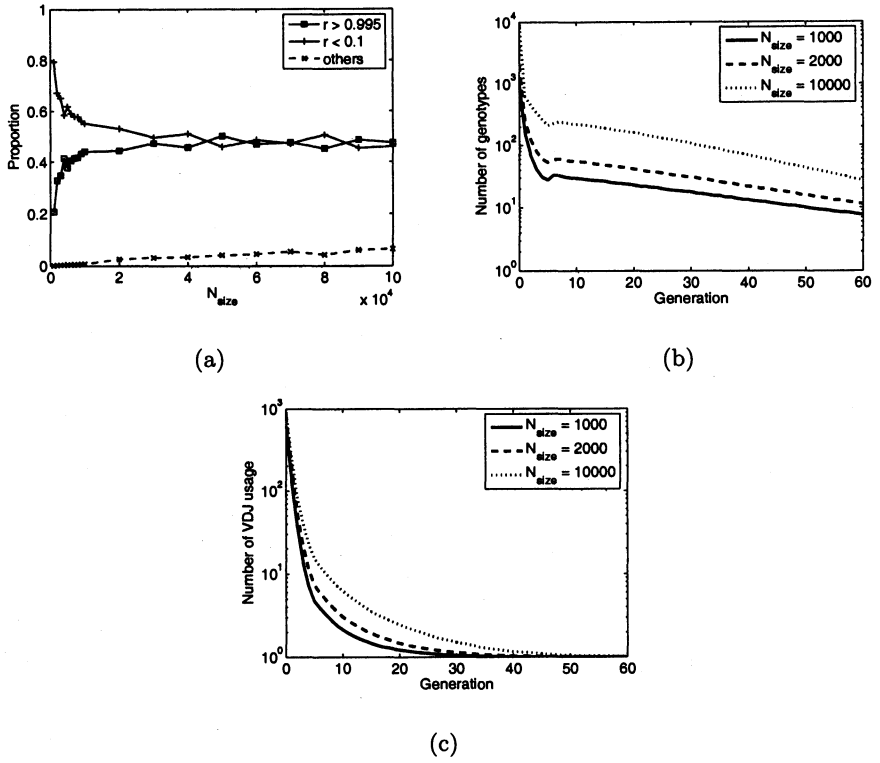


Figure 9.4 : (a) Relationship between the number of B cells reacting to one antigen, defined as N_{size}, and the correlation data between two zebrafish. Measured by the generalized NK model, the correlation coefficient r between the VDJ usage in the mature B cells from the secondary immune response fell into three categories: identical ($r > 0.995$), distinct ($r < 0.1$), and unfixed VDJ usage, with probabilities p, q, and $1 - p - q$, respectively. The values of p, q, and $1 - p - q$ were plotted respectively as the functions of N_{size} ranging from 10^3 to 10^5. (b) The number of distinct genotypes in each generation of the B cells in primary immune response. The x- and y- axes are the same as those in figure 9.2(a). This diagram presents the dynamics of genotype numbers in three cases, $N_{\text{size}} = 1000$, $N_{\text{size}} = 2000$, and $N_{\text{size}} = 10000$, respectively. (c) Same as (b), except for the number of different VDJ recombinations.

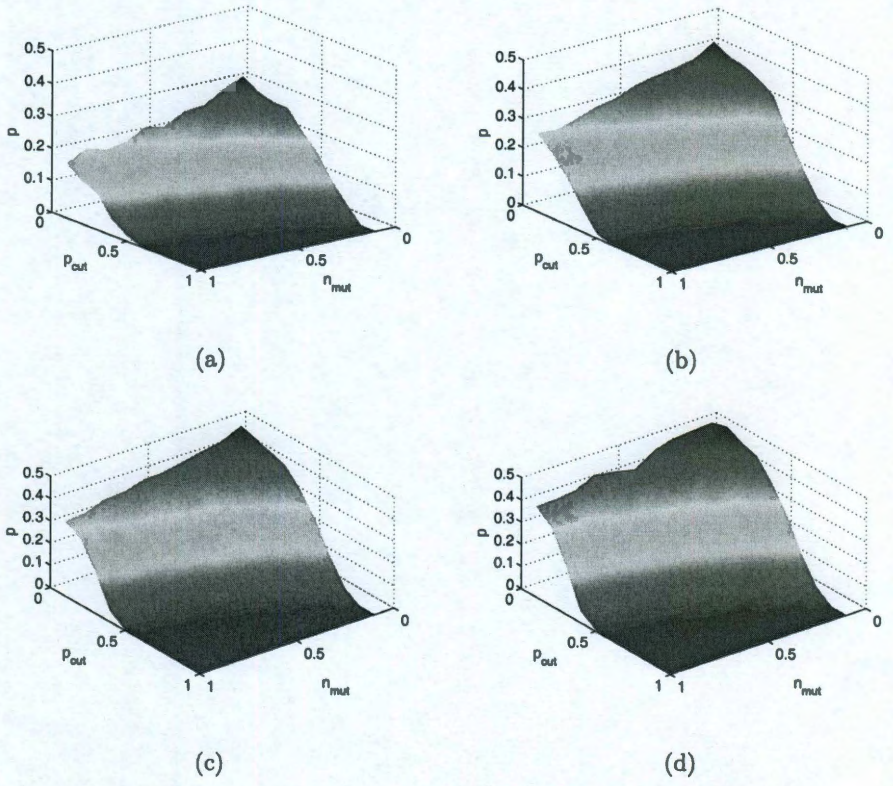


Figure 9.5 : The effect of the average number of point mutations in each generation, n_{mut}, and the proportion of B cells propagated to the next generation, p_{cut}, on the probability p that two zebrafish developed mature B cells with correlated VDJ recombination against the antigen recognized by both zebrafish at the end of the secondary immune response. Each point on the surfaces shows the value of p calculated from the generalized NK model as a function of n_{mut} and p_{cut}. Each of the four subfigures is shown for distinct numbers of antigen-specific B cells N_{size}: (a) $N_{\text{size}} = 1000$, (b) $N_{\text{size}} = 2000$, (c) $N_{\text{size}} = 5000$, and (d) $N_{\text{size}} = 10000$.

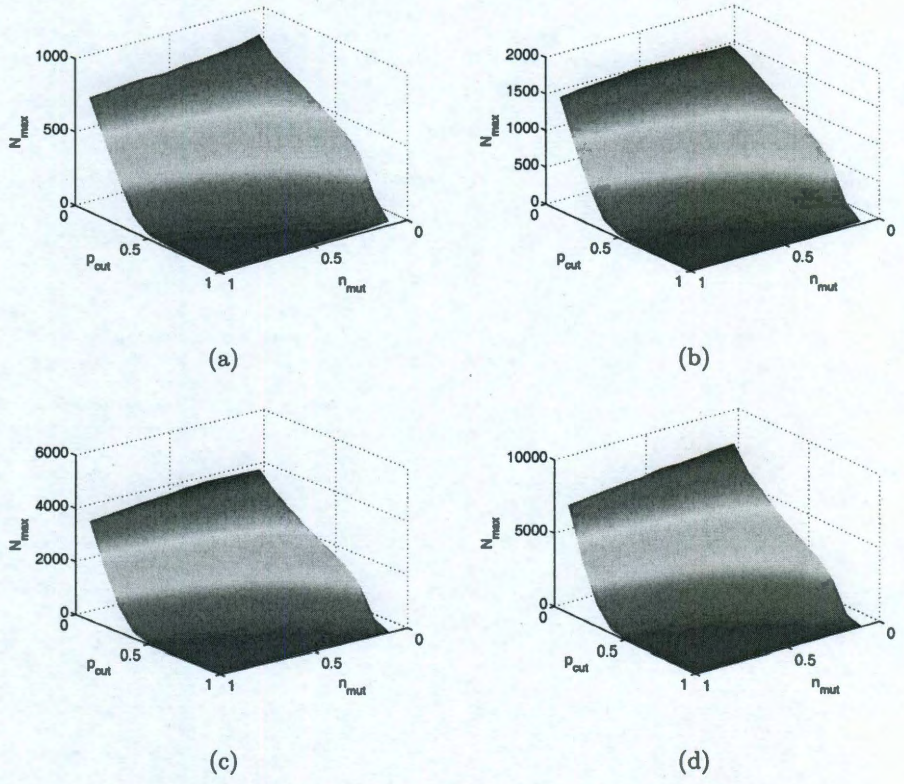
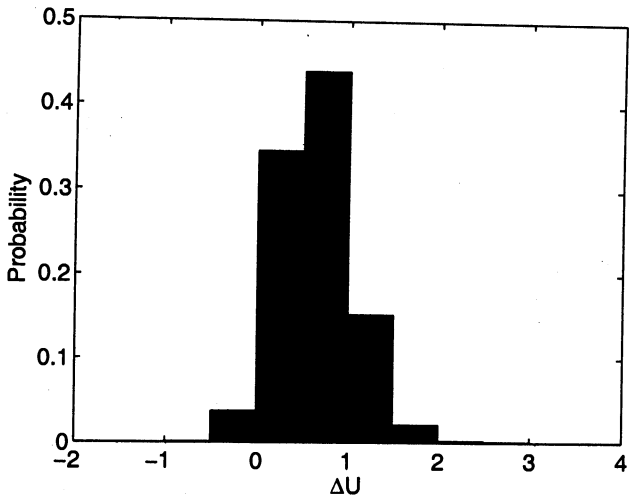
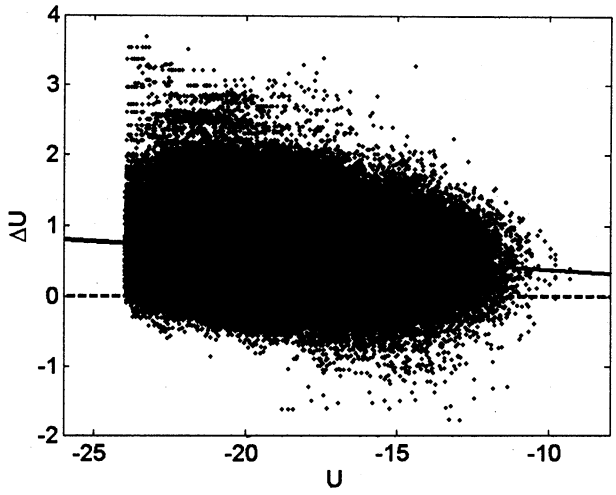


Figure 9.6 : The maximum number of B cells with identical sequence at the end of the secondary immune response, $N_{\max}$, as a function of n_{mut} and p_{cut}. As in figure 9.5, N_{size} as the number of antigen-specific B cells has a constant value in each subfigure: (a) $N_{\text{size}} = 1000$, (b) $N_{\text{size}} = 2000$, (c) $N_{\text{size}} = 5000$, and (d) $N_{\text{size}} = 10000$.



(a)



(b)

Figure 9.7 : (a) The histogram of the energy difference $\Delta U = \hat{U} - U$ associated with a point mutation in the generalized NK model, in which U and $\hat{U}$ are the energy values before and after the point mutation. We calculate U and $\hat{U}$ for 1000 B cells at 61 generations. The values of ΔU ranged from -1.78 to 3.69. The histogram was equally divided in the interval $(-2, 4)$ by 12 bins. The relative frequency of the mutations with $\Delta U < 0$ is 0.038. (b) The original energy, U, versus the energy difference, ΔU, during 20 runs of the generalized NK model. The horizontal dashed line is $\Delta U = 0$. The solid line is the trend line between U and ΔU fit through the data. On average U decreases with the generation. At generation 0, U is typically close to -10, and at generation 60, U is typically close to -25.

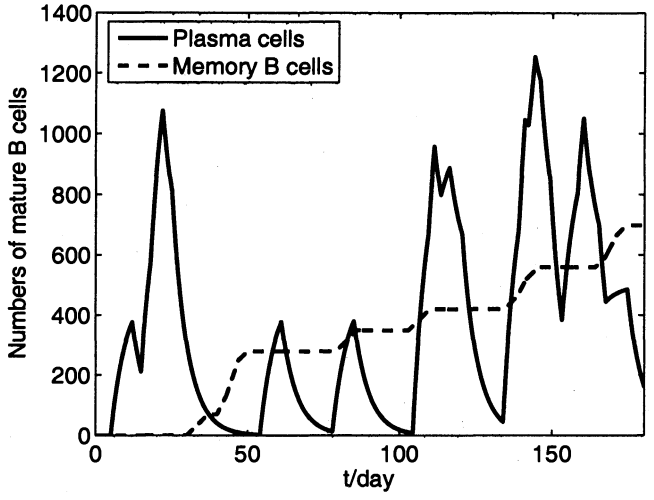
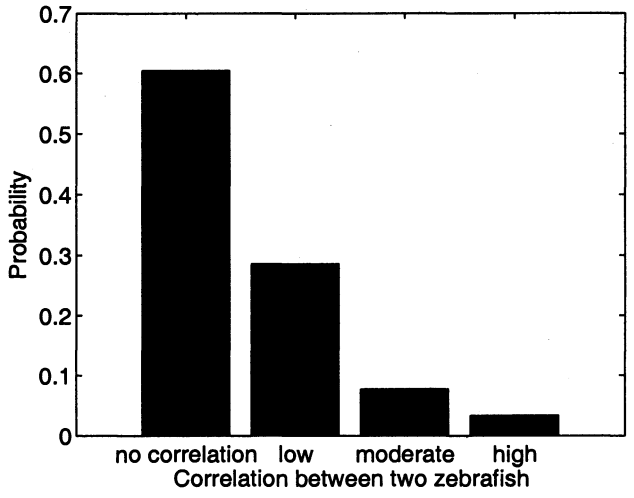
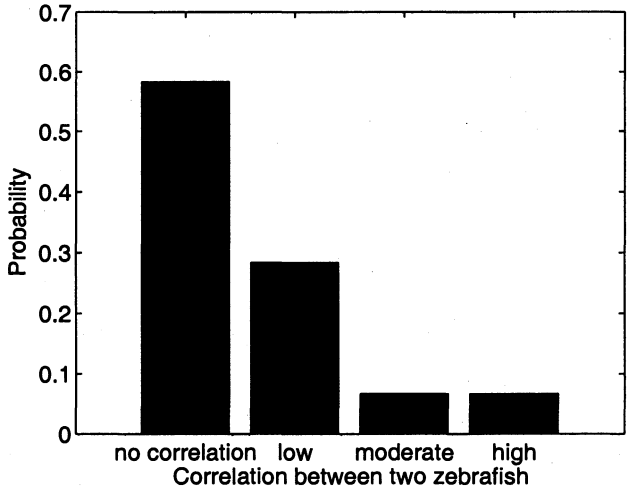


Figure 9.8 : Dynamics of mature B cells on day 0–180 in one zebrafish reacting to all types of the antigen in the environment. The numbers of plasma cells and memory B cells are the sums of all types of plasma cells and memory B cells, respectively. Each zebrafish was challenged by 10 types of antigen, the inoculation time of which followed a Poisson process as described in the main text. The dynamics shown in this figure are from one trajectory of the Poisson challenge process.



(a)



(b)

Figure 9.9 : Distribution of correlation coefficients between the VDJ usage in B cell repertoires in distinct zebrafish on day 180. Using the categorization scheme in [7], the correlation coefficients r fall into four bins: no correlation with $r < 0.1$, low correlation with $0.1 \leq r < 0.2$, moderate correlation with $0.2 \leq r < 0.5$, and high correlation with $r \leq 0.5$. The height of each bar quantifies the relative frequencies of the correlation coefficient data in each bin. (a) Correlation coefficient data in the experiment [7]. (b) A total of 6000 correlation coefficients were generated by the model.

Chapter 10

Understanding Original Antigenic Sin in Influenza with a Dynamical System

Original antigenic sin is the phenomenon in which prior exposure to an antigen leads to a subsequent suboptimal immune response to a related antigen. Immune memory normally allows for an improved and rapid response to antigens previously seen and is the mechanism by which vaccination works. We here develop a dynamical system model of the mechanism of original antigenic sin in influenza, clarifying and explaining the detailed spin-glass treatment of original antigenic sin [34]. The dynamical system describes the virus load as it propagates through healthy and infected cells, the naïve and memory B cell concentrations, and the affinity of the immune response. Explicit correspondences between the microscopic variables of the spin-glass model and the dynamical system model will be given. The dynamical system model reproduces the phenomenon of original antigenic sin, and describes how competition between different B-cells compromises the overall effect of the immune system. The trade off between the naïve and memory immune responses as a function of antigenic distance between the initial and subsequent antigens is displayed. The suboptimal immune response of the original antigenic sin is observed for intermediate antigenic distances.

10.1 Introduction

Immune memory is mounted from previous infection or vaccination, stores the information for recognizing the corresponding antigen, and is activated during the future

infection of the same type of pathogen. Long-term immune memory has been observed for various pathogens including smallpox [89], malaria [90], hepatitis B [91], dengue [92], and Influenza A [93]. This long-lasting effect prevents the reinfection by pathogens such as smallpox via recognizing and rapidly eliminating those reinfected pathogen particles. Smallpox virus, also called variola virus, propagates only in humans and has a relatively low mutation rate [35]. In contrast, Influenza A virus propagates in human, pigs, and aquatic birds, with a higher mutation rate that is approximately 2.0×10^{-6}/nucleotide/infectious cycle [37], or 1.6×10^{-5}/amino acid/day. Calculation of the binding free energy between human antibodies and circulating Influenza A strains shows that the virus mutates away from the genotypes that code for hemagglutinin proteins that are well recognized by the human immune system [94]. Thus for Influenza A, there is a significant antigenic distance between the circulating strain in a given year and the immune memory from previous years.

Original antigenic sin is the phenomenon in which prior exposure to an antigen leads to a subsequent suboptimal immune response to a related antigen. In some years, when the antigenic distance between the vaccine strain and the circulating strains fell into a certain range, the existence of original antigenic sin frustrated the effort to decrease Influenza A infection rate by vaccination. Historical data of the Influenza A vaccine indicates that vaccine efficacy does not monotonically decrease with the distance between the vaccine strains and the circulating strains, but rather has a minimum at an intermediate level of antigenic distance [2]. Interestingly, since the efficacy of the vaccine with this intermediate antigenic distance from the circulating strains is lower than the case with larger distance, which is equivalent to unvaccinated people, original antigenic sin could make vaccinated people more susceptible to the virus than those who are unvaccinated.

One of the earliest works discussing the mechanism of original antigenic sin at the antibody level includes [34], which attributed original antigenic sin to the localization of subsequent immune response in the amino acid sequence space around the primary one. The affinity between an antibody and an antigen is given by the generalized NK model (GNK model) derived from the NK model originally introduced to model a rugged fitness landscape [82, 95] and evolution processes [96, 97, 98], to model the three-dimensional structure of protein molecules rather than peptides. The sequences of a group of antibodies for Influenza A are allowed to mutate freely and independently in the affinity landscape to maximize the individual affinities to the virus. B cells that make antibodies with highest affinities are expanded and propagated to the next round of the simulation. The mutation of the virus is captured by changing the fitness landscape. The final average affinity correlates well with the observed data in history [2].

In this chapter, a set of ordinary differential equations (ODE) is employed to present a deterministic explanation of original antigenic sin equivalent to both the observed data [2] and the GNK model [34]. Previously, various ODE models were established to describe and simulate the process of virus infection and the reaction of the host immune system [239, 240, 241, 242, 243, 244], the basic elements of which are described in [245]. Automatons have been used to model the spatial distribution varying with time of the tissue cells and the virus [246, 247].

The main purpose of this chapter is to build a deterministic model equivalent to the GNK model [34], and to reproduce the observed original antigenic sin phenomenon using an ODE-based deterministic approach. Most of the parameters come from experimental data, leaving a minimal number of parameters to be estimated. The terms in the ODE system have clear physical meanings, so our model explicitly

illustrates the details of the infection process.

10.2 Materials and Methods

10.2.1 Characters of Influenza A Virus and Infection

In the following sections of this chapter, we select Influenza A as the model to discuss the mechanism of original antigenic sin. Influenza A is a type of RNA virus, belonging to the family Orthomyxoviridae. A functional Influenza A particle consists of a spherical lipid bilayer shell, with 8 distinct RNA strains that encode 11 kinds of proteins. Two kinds of glycoproteins, hemagglutinin (HA) and neuraminidase (NA), adhere to the surface of the virus particle. HA contributes to the binding of the virus particle to the sialic acid on the surface of the target upper respiratory tract epithelial cells and facilitates subsequent fusion and entry [18, 19, 20]. NA is the key component that facilitates virus release from surface membrane of infected cells [21]. Nucleocapsid protein (NP), a nucleoprotein, encapsulates and transports viral RNA inside the host cell [22]. The matrix protein, M1, binds to the inner side of the viral lipid bilayer membrane and helps assemble virus particles inside host cells, and is the central component in budding of virus particles [21]. M2, a glycoprotein inside the lipid bilayer, serves as the ion channel adjusting the pH value inside the virus particle, uncoating the virus particle, and contributes to efficient virion replication [23]. Additionally, there are two non-structural proteins, namely NS1 and nuclear export protein (NEP, formerly named NS2), as well as four RNA polymerases replicating the virus RNA in the nucleus of the infected cell, namely PA, PB1, recently discovered PB1-F2 [24], and PB2. NS1 interferes with the cellular antivirus system [25], and NEP exports newly synthesized viral ribonucleoprotein (RNP) complexes comprising

viral RNA, NP, and PB1 from the nucleus [26].

Influenza A infection follows a common dynamical process. The infection occurs in the epithelial cells on the surface of upper respiratory tract in those bronchi that are larger than 3.3 mm [240]. The incubation period between the infection and the emergence of symptoms ranges from one day to five days, but is typically two days. The host starts to shed infectious virus particle approximately 24 hours prior to formation of symptoms. Initially, the typical concentration of Influenza A virus particles is 10^{-13} M. The virus load usually reaches a maximum 3×10^{-9} M in two days after infection, and falls back to the initial level six days after infection [240]. Influenza A virus is cytopathic and destroys the infected cell, causing the dead cells to accumulate *in situ* until they are cleaned. The percentage of dead epithelial cells reaches the maximum percentage of 30–50% on Day 2, and decreases to 10% on Day 5. If the maximum percentage is lower than 10% during the whole process, no symptom will occur [240]. The immune system is activated by the existence of the virus particles. Antibodies IgG and IgA are the most important immune components controlling the disease, followed by the CD8 cytotoxic T-cells [72]. The level of B cells and plasma cells increase by 10^2 times and 2×10^4 times in 7 days, respectively [240]. Primary infection without previous immune memory generates memory antibodies making up 0.1% – 1% of the total antibodies, with the affinity 10^6 M^{-1}, concentration 10^{-13} M [72, 244]. Among the naïve antibodies, those capable to recognize the antigen occupy 0.001% – 0.01% of the total antibodies, with the affinity 10^4 M^{-1}, concentration 10^{-15} M [72, 244].

10.2.2 Model Development and Description

We use a simplified model consisting of the major factors in the tissue and immune system to describe the dynamics of Influenza A infection and recovery, with all the state variables listed in Table 10.1. The concentration of epithelial cells on the upper respiratory tract is kept around a certain homeostatic level H_0, which is also the sum of concentrations of healthy cells (H), infected cells (I), and dead cells killed by the Influenza A virus (D), respectively. Free Influenza A virus particles (V) are released from infected cell, and eliminated by the host immune system. We only consider the effects of antibodies since antibodies are the dominant factor in neutralizing Influenza A virus. Two types of antibodies exist in the model: the naïve antibodies without previous maturation, and the memory antibodies generated and reserved from the last infection of Influenza A. The concentrations of them are defined as X_1 and X_2, respectively. The humoral immune system recognizes and removes antigens bound by antibodies. With the definition of antibody affinity

$$K_{\mathrm{a}} = \frac{[\mathrm{Ag} : \mathrm{Ab}]}{[\mathrm{Ag}]\,[\mathrm{Ab}]} \tag{10.1}$$

the concentration of Influenza A virus particles bound to antibodies is proportional to the concentration of free Influenza A virus particles and antibodies, along with the affinity U_1 and U_2 between the antigens and the naïve and memory antibodies, respectively. The maximal affinity is U_{max}.

From the above information, a set of ODEs is built to describe the scenario that there exist in host immune system naïve antibodies with low initial affinity and memory antibodies with higher initial affinity having persisted since the last infection. A minimal state variable set $\mathbf{Z} = (H, I, V, X_1, X_2, U_1)$ is selected comprising the healthy and infected cells H and I, virus load V, concentrations of naïve antibodies and

memory antibodies X_1 and X_2, as well as the affinity of naïve antibodies U_1.

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \lambda D - \beta V H \tag{10.2}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta V H - aI \tag{10.3}$$

$$\frac{\mathrm{d}V}{\mathrm{d}t} = kI - \mu V - p\, U_1 X_1 V - p\, U_2 X_2 V \tag{10.4}$$

$$\frac{\mathrm{d}X_1}{\mathrm{d}t} = c(\mathbf{Z}, t)\, \frac{U_1 X_1}{U_1 X_1 + U_2 X_2} - bX_1 \tag{10.5}$$

$$\frac{\mathrm{d}X_2}{\mathrm{d}t} = c(\mathbf{Z}, t)\, \frac{U_2 X_2}{U_1 X_1 + U_2 X_2} - bX_2 \tag{10.6}$$

$$\frac{\mathrm{d}U_1}{\mathrm{d}t} = s\, U_1 X_1 V\, (U_{\max} - U_1)\,. \tag{10.7}$$

The homeostasis of epithelial cells gives an additional algebraic function for the concentration of dead cell D

$$D = H_0 - H - I. \tag{10.8}$$

Equation 10.2 describes the dynamics of the concentration of healthy epithelial cells. The repair mechanism is activated only if any damage in epithelial cells is detected ($D \neq 0$), and new healthy cells are regenerated with the rate λD due to this loss [240]. Alternative expressions include regeneration rate $= \lambda$ [243] or λDH [244]. In the stochastic model in [247], the regeneration rate is 0 when $D = 0$, and has a mathematical expectancy of λH when $D \neq 0$. The average life span of trachea cells is 47.5 days in human [248], while the time for epithelial cell regeneration is 0.3 – 1 day [249, 240], showing that the cell division is not the major driving force for cell regeneration. With this consideration in mind, we select the format in [240]. The infection rate β represents affinity between virus and sialic acid, as well as the number of sialic acid molecules on the surface of the cell. The protective effect of interferon is neglected for this simplified model.

Equation 10.3 characterizes the concentration of infected cells. All the infected cells are originally healthy cell, and are killed by the virus by the rate a.

Equation 10.4 depicts the generation and elimination of the free virus particles. Free virus particles are released from the infected cells by the rate k. The half-life of free virus particles is $1/\mu$. Virus bound by antibodies is neutralized, and is the target for the immune system that clears the virus. Thus the clearing rate is proportional to the concentration of antigens bound by antibodies, [Ag:Ab]. From equation 10.1, the clean rate is proportional to the affinity of the antibody U_i, the concentration of corresponding antibodies X_i and free virus particles V, respectively.

Equation 10.5 and 10.6 show the secretion and decay of naïve antibodies and memory antibodies. Antigen presenting cells (APC) process the material from pathogen and present the antigen on their surface, activating naïve T cells. Parts of those activated T cells proliferate and differentiate into Th2 helper T cells. Th2 cells and free virions activate B cells together [72]. The intensity of activating signal for antibodies, $c(\mathbf{Z}, t)$, is a function of time to be determined later, depending on the virus load, APC, and naïve T cells. Naïve B cells mature in germinal centers, undergoing proliferation and somatic hypermutation. Mutated B cells are selected by competing for binding antigen and activation signal from Th2 cells surrounding the germinal center. The morphology of germinal centers determines that the interface between the B cell region and Th2 cell region is approximately constant, and so is the amount of antigens inside the germinal center. Therefore B cells inside the germinal center compete with each other for limited activating signal. The ratio of the intensity of activating signal for naïve and memory B cells is U_1X_1/U_2X_2. The decay rate of both naïve and memory antibodies is b.

Since the hypermutated B cells are selected by the affinity to the antigen, increase

Table 10.1 : Descriptions and units of the variables in the model.

Variable	Description	Unit
H	Concentration of healthy cells	1.7×10^{-11} M
I	Concentration of infected cells	1.7×10^{-11} M
V	Concentration of virus particles	1.7×10^{-11} M
X_1	Concentration of naïve antibodies recognizing the virus	1.7×10^{-11} M
X_2	Concentration of memory antibodies reserved from the last infection	1.7×10^{-11} M
U_1	Affinity of naïve antibodies recognizing the virus	1.0×10^{7} M^{-1}
U_2	Affinity of memory antibodies recognizing the virus	1.0×10^{7} M^{-1}
D	Concentration of dead cells	1.7×10^{-11} M

of the affinity of naïve antibody U_1 is driven by the successful binding between the naïve antibody and the antigen. Equation 10.7 indicates that the increase rate of the affinity is proportional to the concentration of Ag:Ab complex. The logistic factor $(U_{\max} - U_1)$ ensures that the probability for B cells to mutate to a state with higher affinity decreases with the process of maturation.

10.2.3 Reduced Units and Parameter Estimation

To facilitate the numerical calculation, reduced units are used for all the state variables. For the state variables H, I, V, X_1, and X_2, the unit is defined as the homeostatic concentration of epithelial cells in the upper respiratory tract, which is 1.7×10^{-11} M [240, 244]. The unit of U_1 and U_2 is defined as the maximum affinity between memory antibody and antigen, which is 1.0×10^{7} M^{-1} [72]. The reduced units for all the variables in equation 10.2 – 10.7 are listed in Table 10.1.

The majority of parameters in this model are acquired from previous experiments. Those data fall into two groups of numbers that are compatible. A detailed picture of influenza disease in the cellular level was depicted by [240, 239, 244, 249], taking

Table 10.2 : Parameters extracted from experimental data.

	Physical meaning	Parameter set 1	Parameter set 2	Initial estimation
λ	Regeneration rate of healthy epithelial cells	2		
β	Infection rate	0.34	0.27	
a	Death rate of infected epithelial cells	1.5	4.0	
k	Number of virus released by each infected epithelial cell	510	480	
μ	Nonspecific virus removal rate	1.7	3.0	
p	Virus removal rate by antibodies	619.2		
b	Decay rate of antibodies	0.043		
c_0	Production rate of antibodies			1.0
s	Antibody maturation rate			100

into account the concentrations of cells and virions, as well as the APC, interferon, Th1 and Th2 helper cells, CTL, interferon, B cells, plasma cells, and antibodies. This picture gives the first set of parameters. The second set of parameters is extracted from an independent Influenza A infection experiment in 6 human volunteers [242], where a simpler ODE model with a fixed parametric form is built to fit the measured daily virus load from nasal wash. The two sets of parameter presented in the reduced units are listed in Table 10.2. Despite the different approaches, the variables β, a, k, and μ from [240, 239, 244, 249] and [242] are similar.

Compared to some previous models [240, 244], a major simplification in this chapter is neglecting the propagation of the activation signal originated by the detection of virus, through a chain consisting of APC, Th2 cells, and B cells. Instead, we introduce a time-dependent parameter $c(\mathbf{Z}, t)$ as the activation signal for antibodies. In a typical infection process, the level of APC reaches the peak simultaneously with the virus load on Day 2 [240]. Virus load falls back to the initial level on Day 6 [240], while the half lives of APC, helper T cells, B cells, and plasma cells (Table 10.3)

Table 10.3 : Decay rates of different immune cells.

Immune cell	Decay rate/day^{-1}	Reference
APC (in stimulated state)	1	[244]
Macrophage	1	[240]
Th1 helper cell	1	[240]
Th2 helper cell	1	[240]
B cell	0.1	[240]
Plasma cell	0.4	[240]

are comparable to the duration of the whole infection process. Thus the duration of maturation of B cells and the generation of antibodies is estimated to be 14 days, continuing after most Influenza A virus particles are removed. Accordingly, the function $c(\mathbf{Z}, t)$ has the initial value of zero, is assigned the value of c_0 when V reaches 0.1, and remains c_0 for 14 days, before going back to 0 again. From the output of the models in [240, 244], the initial estimation for c_0 is fixed as 1.0, and s is initially estimated as 100.

10.3 Results

10.3.1 Time Courses of Infection and Recovery

With all parameters including c_0 and s defined and fixed, we use the stiff differential equation solver ode23s in MATLAB to numerically solve equations 10.2 – 10.7. The first set of parameters listed in the left column of Table 10.2 is adopted. As described in the Materials and Methods section, at the moment of infection, all the

epithelial cells are healthy cells, the initial virus load is variable, typically 1% of the concentration of epithelial cells. The concentration of naïve antibody capable to recognize the antigen is approximately 10^{-4} of the concentration of epithelial cells, while that of memory antibody is 10^{-2} of the concentration of epithelial cells. Initial affinity of naïve antibodies to the antigen is 10^4 M^{-1}, 10^{-3} of the maximum affinity. Using the previously introduced reduced units, the initial values $\mathbf{Z}(0) = (H(0), I(0), V(0), X_1(0), X_2(0), S_1(0)) = (1, 0, 0.01, 10^{-4}, 10^{-2}, 10^{-4}, 10^{-3})$. The time span for the simulation is 0 – 20 days. The solved trajectories of state variables $\mathbf{Z}$ are compared to the kinetics observed in reality to verify the values of parameters.

Two cases are simulated to describe the dynamics of all state variables. The first one has a weak cross immunity between memory antibodies generated in the previous infection and the current virus, corresponding to the scenario that the virus mutates substantially from the previous strains. The second one has a strong cross immunity, which is used to simulate the infection caused by Influenza A without intensive escape mutation. The affinity of memory antibodies is $U_2 = 10^{-3}$ for the first case, and $U_2 = 0.5$ for the second case. The details of the dynamics are shown in Figure 10.1 and 10.2.

Figure 10.1 describes the whole process of infection and recovery with a weak immune memory $U_2 = 10^{-3}$. Symptoms with approximately 30% of the epithelial cells killed is observed at the beginning of the infection. The peak of the proportion of dead cell D occurs on Day 1 – Day 2, agreeing with the experimental data. On Day 5, the percentage of dead cell falls under 10%. At the same time of the increase of the percentage of dead cells, the virus load V increases by $10^3 - 10^4$ fold, reaching the climax on Day 1, similar to the experimental results of $10^4 - 10^5$ fold increase on Day 2. The virus load V decreases to the initial level on Day 3 – Day 4. A 10^5 fold increase

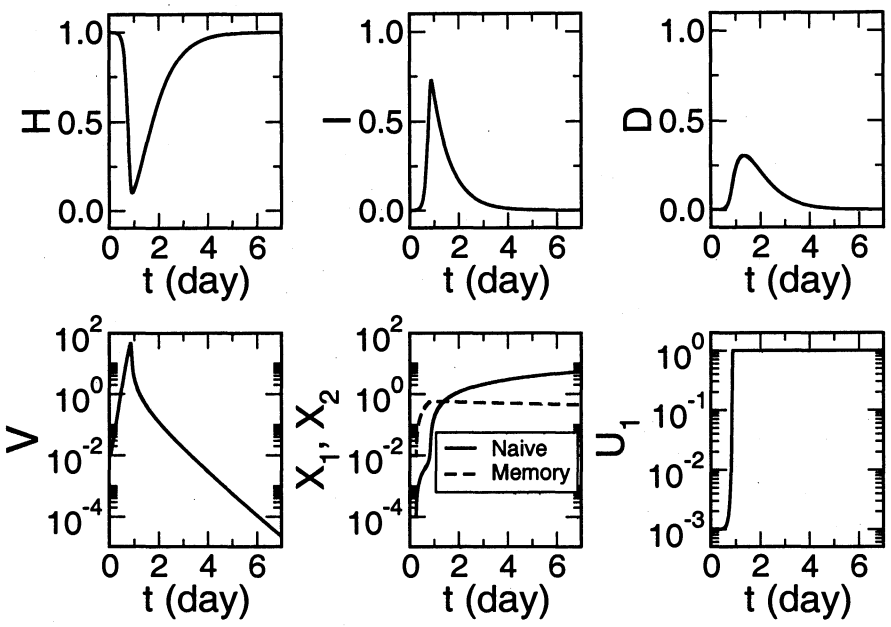


Figure 10.1 : Time courses of proportions of healthy cell (H), infected cell (I), and dead cell (D), virus load (V), concentration of naïve and memory antibodies (X_1 and X_2), and the affinity of naïve antibody (U_1). with the condition $U_2 = 10^{-3}$. Initially, $H(0) = 1$, $I(0) = D(0) = 0$, $V(0) = 0.01$, $X_1(0) = 10^{-4}$, $X_2(0) = 10^{-2}$, and $U_1(0) = 10^{-3}$.

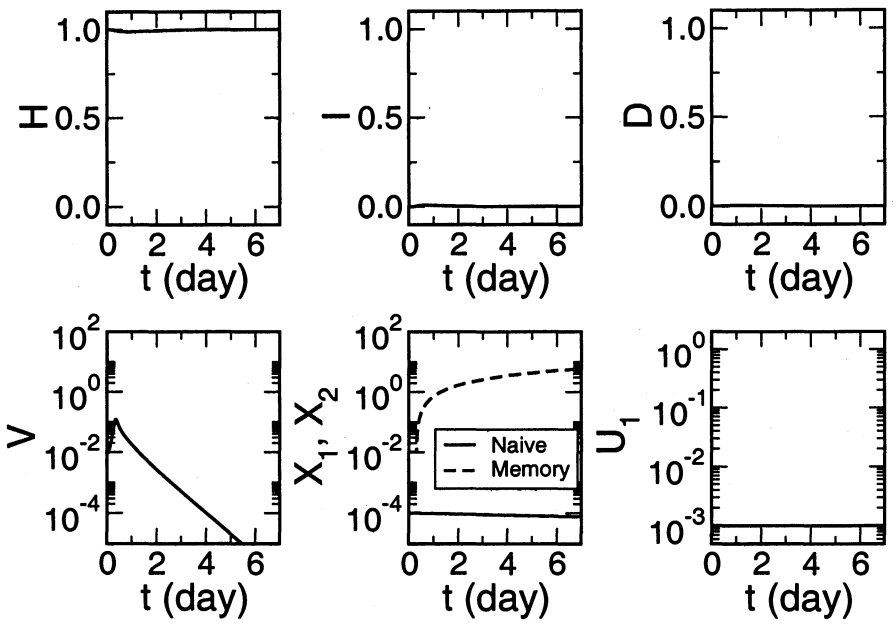


Figure 10.2 : Time courses of proportions of healthy cell (H), infected cell (I), and dead cell (D), virus load (V), concentration of naïve and memory antibodies (X_1 and X_2), and the affinity of naïve antibody (U_1). We set the condition $U_2 = 0.5$. The initial values of all the state variables equal those in Figure 10.1.

of the concentration of naïve antibody occurs during the infection and recovery, and the affinity of naïve antibody approaches to the maximum level $U_{\max} = 1$. The concentration of memory antibodies X_2 has an initial 10^2 fold increase, and decreases approximately exponentially after Day 1 with the rate 0.0427, similar to the decay rate of antibodies $b = 0.043$. Thus almost no memory antibody is produced after Day 1.

Figure 10.2 shows the dynamics with a strong immune memory $U_2 = 0.5$. No obvious proportion of dead cells is accumulated and thus no symptoms are observable in the infected person. The virus load is suppressed remarkably compared to Figure 10.1. This scenario depicts the effect of a successful vaccination. Compared to Figure 10.1, the increase of the concentration of naïve antibodies disappears, the value of X_1 decreases approximately exponentially with the rate 0.0423, close to the decay rate of antibodies $b = 0.043$, indicating that naïve antibodies are barely produced during the whole process. No significant somatic hypermutation is observed in those corresponding B cells, hence the naïve antibody affinity almost keeps constant, which is confirmed in the plot of U_1. There is a larger expansion of the memory antibodies: the value of X_2 on Day 7 is approximately 10 fold higher than the corresponding value in Figure 10.1.

Naïve antibodies dominate the memory antibodies with lower affinity, and memory antibodies with high affinity dominate naïve antibody. The transition of these two types of antibodies corresponds to a critical region of the memory antibody affinity U_2. In the following discussion, values of U_2 in the whole range $10^{-3} \leq U_2 \leq 1$ are examined in the model, rendering a picture of the changing character of the dynamics, such as the peak of virus load and dead cell percentage, as well as the cumulative effect and the average affinity of antibodies. The model will reproduce original antigenic

sin shown in the experimental data in the intermediate level of U_2.

10.3.2 A General Picture of Original Antigenic Sin

To illustrate the phenomenon of original antigenic sin, values of U_2 in the range 10^{-3} to 1.0 are chosen. The minimum value, $U_2^{\min} = 10^{-3}$, reflects the case that the previously generated memory antibodies barely recognize the antigen, and the maximum value, $U_2^{\min} = 1.0$, demonstrates the strongest memory immunity with largest affinity of the antibodies. The intermediate level of U_2 simulates the scenario in which previously generated antibodies have the decreased capability to recognize the antigen due to the escape mutation of the Influenza A virus or imperfect vaccine selection. 100 independent simulations were run for all these cases. The maximum virus load and the maximum percentage of dead cells were recorded for each simulation. The integrated effects of naïve antibodies and memory antibodies are calculated with

$$X_1^{\text{Int}} = \int X_1(t)\, U_1(t)\, \mathrm{d}t \tag{10.9}$$

$$X_2^{\text{Int}} = \int X_2(t)\, U_2(t)\, \mathrm{d}t. \tag{10.10}$$

Similarly, the final average affinity of the antibodies for each simulation is

$$U^{\text{avg}} = \left\{ \frac{X_1(t)\, U_1(t) + X_2(t)\, U_2(t)}{X_1(t) + X_2(t)} \right\} \Bigg|_{t=20}. \tag{10.11}$$

By equation 10.7, the affinity of naïve antibodies (U_1) is monotonically increasing while the affinity of memory antibodies (U_2) is kept constant, so equation 10.11 shows the average antibody affinity after the 20-day maturation of naïve antibodies, when the patient has recovered from the disease. Figure 10.3 depicts the maximum virus load, percentage of dead cells, integrated effects of naïve and memory antibodies, and final average affinity of the antibodies as the function of the affinity of the memory antibodies generated in the previous Influenza A infection.

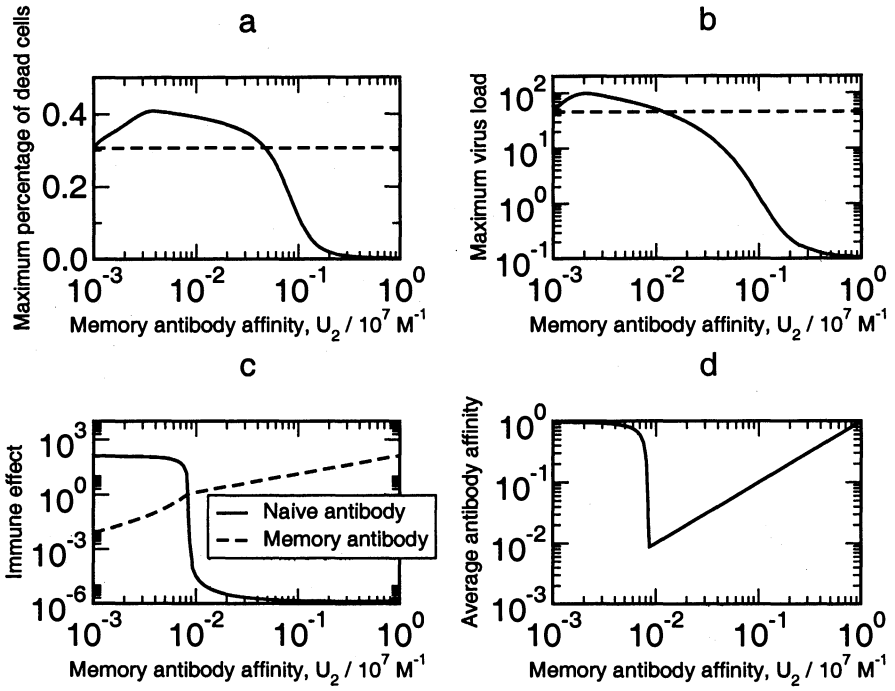


Figure 10.3 : Trajectories of maximum virus loads, maximum percentages of dead cell, maximum immune effects of naïve and memory antibodies by equation 10.9 and 10.10, respectively, and trajectory of final avarage affinities of the antibodies by equation 10.11 in a series of independent simulations, as the function of the affinity of the memory antibodies (U_2). The dashed lines in a and b are the level of maximum virus loads and maximum percentages of dead cells with the lowest memory antibody affinity $U_2 = 10^{-3}$, respectively.

Figure 10.3a presents a picture of original antigenic sin in the viewpoint of maximum percentage of dead cells. With $U_2 = U_2^{\min} = 10^{-3}$, the maximum percentage of dead cell $D_0^{\max}$ is 30.6%. Original antigenic sin is observed in the interval $10^{-3} < U_2 < 5.0 \times 10^{-2}$. The peak of this curve, 40.8%, is obtained with $U_2 = 3.8 \times 10^{-3}$, and is 33.3% higher than the maximum percentage with lowest U_2. This 33.3% increase is significant for the percentage of dead epithelial cells. When $U_2 > 0.1$, the maximum percentage of dead cells falls below 10% indicating no obvious symptom is observed in the infected person. When the antigenic distance from the immune system to the Influenza A virus is small, the immune system can control the virus effectively.

Like Figure 10.3a, Figure 10.3b shows the nonmonotonicity of maximum virus load V during the infection and recovery process as a function of the affinity of memory antibodies. The maximum virus load $V_0^{\max}$ with lowest memory antibody affinity $U_2^{\min} = 10^{-3}$ is $V_0^{\max} = 45.0$. The maximum virus load with $10^{-3} < U_2 < 1.2 \times 10^{-2}$ is higher than $V_0^{\max} = 45.0$, where original antigenic sin occurs. The maximum virus load in the interval $1.6 \times 10^{-3} < U_2 < 2.8 \times 10^{-3}$ doubles $V_0^{\max}$. With $U_2 > 0.11$, the maximum virus load is effectively suppressed to be less than unity, agreeing with the fact that memory antibodies effectively recognize and eliminate pathogens with weak escape mutation [2].

Figure 10.3c describes the cumulative effect X_1^{Int} and X_2^{Int} of both naïve and memory antibodies calculated with equation 10.9 and 10.10 during the whole process of infection. There is a threshold $U_2 = 5.0 \times 10^{-3}$ below which the effect of naïve antibodies is in a plateau with $X_1^{\text{Int}} > 100$. X_1^{Int} decreases sharply from 100 to 9.0×10^{-5} in a narrow region $5.0 \times 10^{-3} < U_2 < 9.3 \times 10^{-3}$. X_2^{Int} increases almost linearly with U_2 from 7.7×10^{-3} ($U_2 = 10^{-3}$) to 133.8 ($U_2 = 1.0$). Between the

cumulative effect of naïve and memory antibodies X_1^{Int} and X_2^{Int}, X_1^{Int} is larger than X_2^{Int} when $U_2 < 8.1 \times 10^{-3}$, and smaller otherwise. Since X_1^{Int} decreases quickly with U_2 near $U_2 = 8.1 \times 10^{-3}$, there is a large difference between X_1^{Int} and X_2^{Int} for most values of U_2, indicating that there is generally one dominant type of antibody, naïve or memory, which makes the major contribution to the control of Influenza A.

Figure 10.3d plots the final average affinity U^{avg} of all the antibodies after recovery. The form of equation 10.7 ensures the monotonic increase of U_1, whereas the trend of increase is highly changeable depending on U_2. Similar with Figure 10.3(c), a plateau with $U^{\text{avg}} > 0.9$ exists with $U_2 < 4.0 \times 10^{-3}$, and U^{avg} decreases substantially from 0.82 to 8.8×10^{-3} when U_2 increases from 5.3×10^{-3} to 8.7×10^{-3}. Note that the sudden decrease in the overall effect of naïve antibodies in Figure 10.3 also occurs in the same region of U_2. When $U_2 > 8.7 \times 10^{-3}$, U^{avg} increases approximately linearly with U_2, which is mainly the contribution of the affinity of memory antibodies.

10.3.3 Mechanism of Original Antigenic Sin

The dynamics system defined by equation 10.2 – 10.7 is split into two subsystems, i.e. an actuator and a controller, with weak coupling between them. Equation 10.2 – 10.4 is the actuator with H, I, V as the state variables, while equation 10.5 – 10.7 is the controller with X_1, X_2, U_1 as the state variables. The control of the actuator is implemented by the expression $E = U_1 X_1 + U_2 X_2 > 0$. That is, equation 10.4 is equivalent to

$$\frac{\mathrm{d}V}{\mathrm{d}t} = kI - \mu V - pEV. \tag{10.12}$$

The actuator consisting of equation 10.2, 10.3, and 10.12 has two steady states:

$$\begin{aligned}(H(\infty), I(\infty), V(\infty))_1 &= (1,0,0)\\ (H(\infty), I(\infty), V(\infty))_2 &= \left(\frac{a(\mu+pE)}{k\beta}, \frac{\lambda[k\beta - a(\mu+pE)]}{k\beta(a+\lambda)}, \frac{\lambda[k\beta - a(\mu+pE)]}{\beta(a+\lambda)(\mu+pE)}\right).\end{aligned}$$

By calculating the eigenvalues of the Jacobian of the actuator, we see the first steady state is stable for any $E > 0$, the second one is stable only if $0 < E < 0.18$. With strong immune response, E is large and the only steady state reflects the complete recovery from the disease. During the process, large E also ameliorates the infection of healthy cells by directly suppressing the proliferation of virus. Numerical simulation for the actuator illustrates the dependence on E of the dynamics of Influenza A infection. Figure 10.4 displays the effect of E: E larger than 0.18 removes all the virus and dead cells, and larger values of E result in higher decay rate of both D and V. Therefore, E is the single factor in this model controlling dead cell proportion and virus load.

The other subsystem is the controller comprising the state variables X_1, X_2, and V. The controller observes the state variables in the actuator in the form of the time-variant factor $c(\mathbf{Z}, t)$, which jumps from 0 to c_0 when the virus load V reaches 0.1, and remains c_0 for 14 days. Due to the quick proliferation of virus in the initial stage of infection, $c(\mathbf{Z}, t)$ approximates the constant c_0. The dynamics of factor E is expressed as

$$\begin{aligned}\frac{dE}{dt} &= \frac{d}{dt}(U_1X_1 + U_2X_2)\\ &= U_2\frac{d}{dt}(X_1 + X_2) + X_1\frac{dU_1}{dt} + (U_1 - U_2)\frac{dX_1}{dt}.\end{aligned} \tag{10.13}$$

The first term in the right hand side of equation 10.13, $U_2 d(X_1 + X_2)/dt$, is the product of U_2 and the derivative of a first order process $X_1 + X_2$, the latter of which is

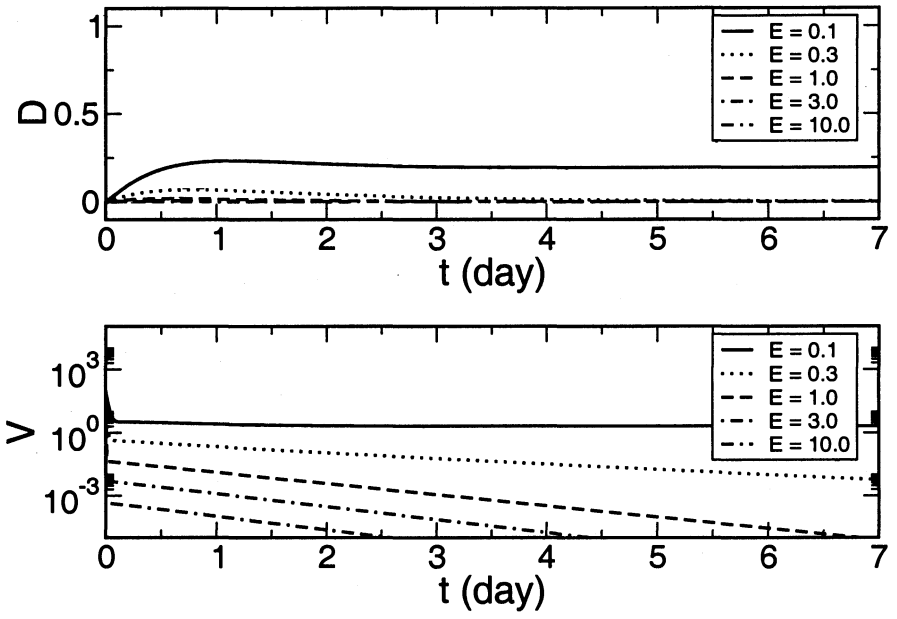


Figure 10.4 : Trajectories of dead cell proportion (D) and virus load (V) with different (E). In each trajectory, $H(0) = 1$, $I(0) = 0$, $V(0) = 100$, and E is constant. Small E such as 0.1 is not able to remove all the virus when $t \to \infty$. Larger decay rates of both D and V are observed for larger E.

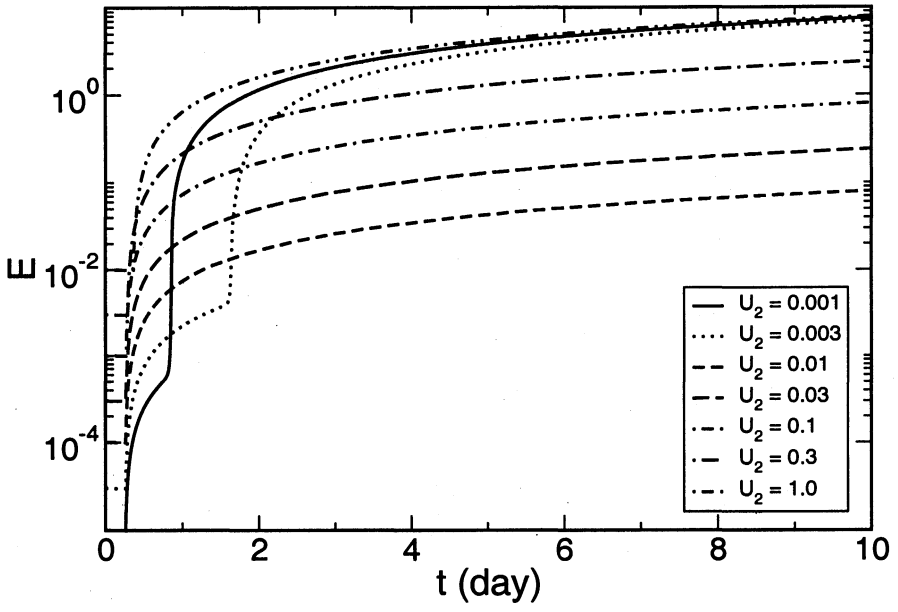


Figure 10.5 : Trajectories of the controlling effect $E = U_1X_1 + U_2X_2$. In each trajectory, $H(0) = 1$, $I(0) = 0$, $V(0) = 0.01, X_1(0) = 10^{-4}, X_2(0) = 10^{-2}$, and $U_1(0) = 10^{-3}$. Each trajectory corresponds to one value of U_2. Small E such as 0.1 is not able to remove all the virus when $t \to \infty$. Larger decay rates of both D and V are observed for larger E.

independent of U_2 by adding equation 10.5 to equation 10.6. The form of equation 10.5 and 10.7 determines the suppression on $X_1(t)$ and $U_1(t)$ by U_2, hence the monotonic decrease of the term $X_1 \mathrm{d}U_1/\mathrm{d}t$ with the increase of U_2. In the case with small U_2, the factor $(U_1 - U_2) > 0$ during almost the whole the process (see Figure 10.1), thus the third term $(U_1 - U_2)\,\mathrm{d}X_1/\mathrm{d}t$ decreases with U_2. If U_2 is large, X_1 is approximately constant, and the change of this term is negligible. Consequently, the first term in equation 10.13 goes up with U_2 and the other terms goes down with U_2. Figure 10.5 shows the quantitative property of E as a functional of U_2: when U_2 increases from 10^{-3} to 1, increase of the second and the third terms of E do not compensate the decrease of the first term, yielding a suppression of E in the intermediate level of U_2.

The source of original antigenic sin is the interaction among the state variables in the controller, or the immune system. When U_2 is small or large, one type of antibody is dominant so that the immune system responds with having only one type of antibody. For intermediate U_2, the interaction and competition of two types of antibodies lead to a decreased immune effect E, which yields less regulation of the tissue. This is original antigenic sin. We observe increased influenza illness rate [2] related to the increase of D and V during the process, as well as decreased final average affinity of antibodies [34] related to the decrease of E. Note that the average affinity is defined as $U^{\text{avg}} = E/(X_1 + X_2)$ where $(X_1 + X_2)$ are approximately independent of U_2 as discussed above.

10.3.4 Sensitivity Analysis

Sensitivity analysis for most of the parameters in this model has been performed in [244]. So we focus on two remaining parameters – c_0 and s. The parameter c_0 characterizes stimulation of the immune system when the virus load increases beyond the threshold 0.1. Since a simplified model comprising the most important factors of both tissue and immune system is given in this chapter, the effects of APC and Th2 cells are combined into the parameter c_0. The parameter s reflects the somatic hypermutation process of the B cell to produce antibodies with high affinity to the antigens. Here we give a sensitivity analysis to each parameter.

Figure 10.6 describes the behavior of the dynamical system with different parameters c_0 and s. The patterns of maximum percentage of dead cells with large U_2 are insensitive to s, and the patterns of average antibody affinity with small and large U_2 are also insensitive to both c_0 and s, but sensitive with intermediate U_2. The threshold where the memory antibodies replace the naïve antibodies as dominant de-

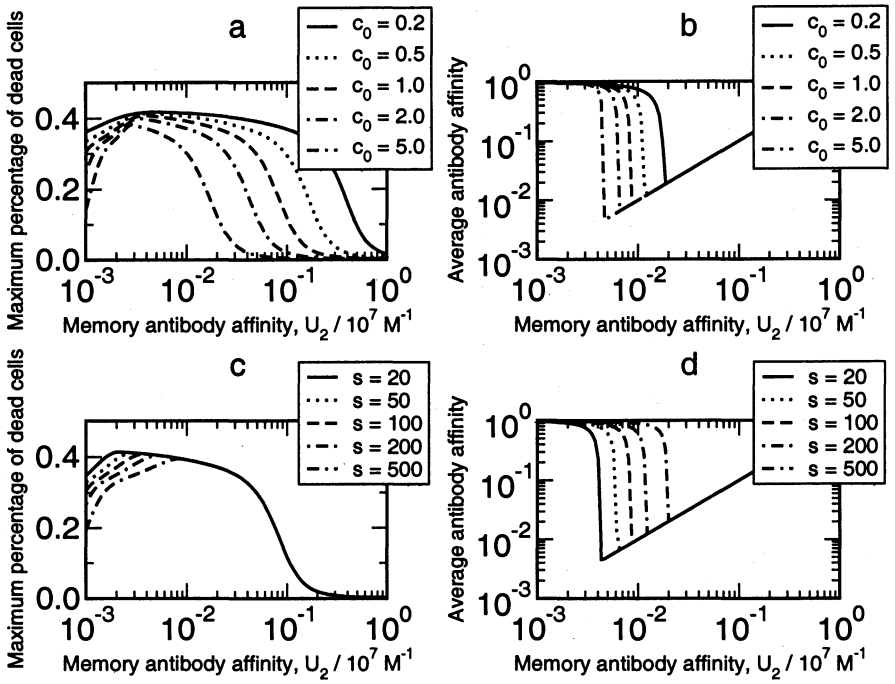


Figure 10.6 : Sensitivity analysis of parameters c_0 and s. (a) and (b) Maximum percentage of dead cells and average antibody affinity for different c_0. (c) and (d) Maximum percentage of dead cells and average antibody affinity for different s. Initial conditions and parameters other than c_0 and s are the same as those in Figure 10.3.

creases with increasing c_0 and increases with increasing s. However, the dynamics for all these values of c_0 and s reproduce the general trend for the common dynamical process of Influenza A infection discussed in Section 10.2.1. Thus the existence of original antigenic sin is also guaranteed for all these parameters sets.

10.4 Discussion

The ODE model expressed in terms of equation 10.2 to 10.7 represents a significant simplification to the previous models describing Influenza A kinetics [240, 244], while introducing a second type of antibodies to capture the competition and cooperation

between antibodies of different genotypes. The effects of Th1 cells, CTL, interferon, and cells protected by interferon are in aggregate by the parameter values of the model. The antibody activation chain consisting of APC, Th2 cells, and B cells is captured by the factor $c(\mathbf{Z}, t)$ rather than explicitly modeled with differential equations.

The concentration of CTL increases by 100 times in the first 7 days after the infection [240] to remove infected cells. The cross immunity is usually strong in cellular immune system for different Influenza A strains [240], while the effect of humoral immune system decreases substantially for a new Influenza A strain with a large evolutionary distance from the previous strain infecting the host [2]. Thus the effects of CTL against different invading Influenza A strains are more homogeneous than those of antibodies. That is, CTL induced by the previous Influenza A infection generally retain the cross immunity against the current invading virus despite virus evolution [240]. In contrast, reaction of antibodies decreases with increasing antigenic distance increases between the current virus strain and the prior exposure strain. Thus the effect of CTL is more stable compared to that of antibody, and can be accounted for by constant parameter values to describe the pattern of original antigenic sin.

For the same reason, the protection for healthy cells by interferon secreted by infected epithelial cells is not explicitly considered either. Additionally, interferon and protected cells are not essential to the dynamics of infection and recovery: absence of interferon does not affect the final elimination of all virus and dead cells [244]. As the elements in the activation chain, APC, Th2 cells, and B cells have little interaction with the elements in the model outside the activation chain, instead, their main function is to activate the downstream elements. Hence a simple function $c(\mathbf{Z}, t)$ is introduced to capture activation.

Equation 10.2 – 10.4 constitute a general form for infection in tissue caused by a cytopathic virus that is controlled mainly by antibodies. This model can be extended to consider the details of CTL and interferon in the immune system. The model could also be extended to consider a family of cytopathic viruses by using equations 10.2 – 10.4 with different parameter sets $(\lambda, \beta, a, k, \mu)$ characterizing each virus in pathogen space. The cytopathic viruses fall into two categories: those inducing acute disease and those inducing chronic disease. Virus in the first category, such as Influenza A virus, are cleared in a short period of time, thus relatively few escape mutations occur. The immune system therefore mutates itself towards a fixed target. The differential equations 10.5 – 10.7 describing the immune system do not require any modification for the escape mutation of the virus. On the other hand, virus in the second category including HIV, persist for years in the host, and keep mutating away from the immune system. A new set of differential equations is needed to model this case. First, additional terms are required for equation 10.7 to describe the trend to decrease the affinity of memory antibodies. Second, if the immune system is also attacked by the virus, equation 10.5 and 10.6 should also contains terms for this effect.

Despite possessing different mathematical forms, there is an implicit parallelism between the dynamical model presented in this chapter and the spin glass model [34]. Both models consider naïve and memory antibodies together, and introduce a competition factor modeling the competition for the survival and opportunity for somatic hypermutation between these two types of antibodies. In the dynamical model, the maturation of the naïve antibody follows a logistic process with a maximum affinity, while in the spin glass model, the antibodies have random walks on a rugged and random landscape [82] where the density of neighboring states with higher affinity is

low for states with high affinity. The dynamical model keeps two antibody strains for naïve and memory affinity, respectively. The spin glass model, however, tracks 1000 naïve antibodies and another 1000 memory antibodies. After a 30-round iteration, the number of antibodies with different genotypes generally converges to fewer than 5, which is close to the dynamical model where both naïve and memory antibodies are monoclonal.

The ODE system applied in this model is deterministic while the real process of virus infection and recovery is a stochastic process. Stochastic factors have been established for a stochastic dynamic system that obtained equivalent results to this model [34]. The dynamical model assumes both naïve and memory antibodies to be monoclonal, which could be changed in the ODE by introducing more model genotypes for both kinds of antibodies. Hence the competition factors $U_1X_1/(U_1X_1 + U_2X_2)$ and $U_2X_2/(U_1X_1 + U_2X_2)$ could be replaced by a tournament-like algorithm involving all the surviving antibodies. With the method of splitting the system into an actuator and a controller, this model also provides a starting point for analysis by nonlinear control theory to explain suboptimality in the immune system. The present simple form, however, provides valuable input to the vaccine development process.

Chapter 11

Conclusion

For both H1N1 (Chapter 2) and H3N2 [2], the HI assay correlates less well with vaccine effectiveness than does p_{epitope}. Collection of HI assay data measuring antigenic distance is also more time-consuming and more expensive compared to the p_{epitope} model. Many hundreds of strains are circulating and collected in an average flu season, thus an HI table with tens of thousands of entries needs to be built to assess the antigenic distance between each pair of strains. With the high-throughput sequencing technology generating hemagglutinin sequence data, such antigenic distances are easily measured with the sequence-based antigenic distance measure p_{epitope}, which correlates to a greater degree with vaccine effectiveness than do the HI data.

The p_{epitope} model is developed to provide researcher and health authorities with a new tool to quantify antigenic distance and design the vaccine. We do not suggest that p_{epitope} should substitute for the current HI assay, but rather suggest that p_{epitope} serves as an additional assessment when selecting vaccine strains. Using p_{epitope} to supplement to HI assay data may allow researchers and health authorities to more precisely quantify the antigenic distance between dominant circulating strains and candidate vaccine strains. The adoption of the p_{epitope} theory may also allow researchers to minimize the cost and the number of ferret experiments and to correct HI assay data in some situations.

The p_{epitope} model is applied to predict vaccine effectiveness and to select vaccine strains. Interestingly, the expected vaccine effectiveness is greater against H1N1 than

H3N2, suggesting a stronger immune response against H1N1 than H3N2. The evolution rate of hemagglutinin in H1N1 is also shown to be greater than that in H3N2, presumably due to greater immune selection pressure.

Shannon entropy is a powerful tool to determine the epitope regions in H1 hemagglutinin. We use Shannon entropy and relative entropy as two state variables of H3N2 evolution. The entropy method is able to predict H3N2 evolution and migration in the next season. First, the Shannon entropy data in one season strongly correlate with the relative entropy data from that season to the next season. If higher Shannon entropy of the virus is observed in one season, higher virus evolutionary rate is expected from this season to the next season. Second, the relative entropy values between virus sequences from China, Japan, the USA, and Europe indicate that the H3N2 virus migration from China to Japan and the USA, and identify a novel migration path from the USA and Europe. The relative entropy values in and out of China, the epicenter, show that evolutionary rate is higher in China than in the migration paths. Moreover, the entropy method was demonstrated on two applications. First, selection pressure of the H3 hemagglutinin is mainly in 54 amino acid positions. Second, the top exposed part in the three-dimensional structure of HA trimer covered by epitopes A and B is under the highest level of selection. These results substantiate current thinking on H3N2 evolution, and show that the selection pressure is focused in a subset of amino acid positions in the epitopes, with epitopes A and B on the top of hemagglutinin dominant making the largest contribution to the H3N2 evolution. These predictions and applications show that the entropy method is not only predictive but also descriptive.

Mutation of the H3 hemagglutinin on the surface of the the influenza virus decreases the ability of the immune system to recognize the flu and decreases the efficacy

of the annual vaccine. We show that influenza tends to increase the number of charged amino acids in the regions of hemagglutinin that the immune system recognizes, probably because this reduces ability of antibodies to bind hemagglutinin. An interesting corollary of this selection is that the number of charges in the dominant epitope of the dominant circulating virus strain is never fewer than that in the vaccine strain, chosen early in the season. We developed a model of the evolution of charge in hemagglutinin by partitioning 20 amino acids into two categories: charged and uncharged, calibrated this model on virus evolution data in humans, and demonstrated the model on Guinea pig animal model studies.

Protein evolution models such as PAM model and BLOSUM model typically apply to the evolution of bacteria, archaea, and eukaryota. For influenza virus, the harsh and changing environment due to immune pressure on the virus, makes its evolution a non-equilibrium dynamics, especially in the time period after its initial emergence in humans. The RAMM model supports the hypothesis that the rate of charge evolution is greater in regions of hemagglutinin recognized by the immune system than in proteins in general. Such temporal and spatial heterogeneity requires a method such as we have presented here for modeling the virus evolution.

We introduced the Einstein crystal as a technology to improve the results of free energy calculation. By calculating the free energy difference of each amino acid substitution, we obtained the free energy landscape for substitutions in epitope B of hemagglutinin. There is notable variation between the values of free energy differences of different substitutions at different sites, because the identities of original and substituting amino acids, as well as the locations of amino acid substitutions, affect to differing degrees the antibody binding process. In this free energy landscape, we suggest that virus tends to evolve to higher $\Delta\Delta G$ values to escape binding of antibody.

Counterbalancing this selection is random drift. Historical amino acid substitutions in epitope B and Monte Carlo simulations of the virus evolution using the free energy based virus fitness, in which random genetic drift of the virus adds statistical noise into the virus evolution process, showed that selected substitutions are biased to those with positive $\Delta\Delta G$ values.

Due to the genetic proximity of the TEM and SHV β-lactamase genes in bacteria with different resistance phenotypes, with typically 1–5 amino acid substitutions conferring resistance to antibiotics, it has been long suspected that transitions of resistance between different antibiotics comes about by accumulation of single DNA point mutations at specific locations. It has been well established that recombination plays a role in the mechanisms that bring about the higher rates of mutation in stressed bacteria. To our knowledge, at the onset of this research there had been none and currently there is one characterization of recombination among naturally evolved β-lactamase resistance variants from clinical strains [197]. We analyzed DNA sequences of the genes for naturally evolved variants of TEM and SHV β-lactamases to establish if recombination plays an evolutionary role in resistance development by combining mutations originating in different genes. All but one of our detection methods indicated the presence of recombination signatures in the gene sequences from clinical isolates. Our findings support a role of recombination in the evolution of extended spectrum antibiotic resistance, suggesting that recombination directly acts on resistance genes to recombine selected mutations.

HA specific immunity plays a key role in H3N2 virus evolution. The fixation of mutations of H3N2 virus in both naïve and immune animal appeared to be non-random events as immune pressure can lead to mutations accumulating in dominant epitopes. We found that the mutations driven by immunity are frequently associated

with charge, hydrophobicity, and lost glycosylations. The results of the present study increase our understanding of the direction of evolution in influenza and may provide useful insights for the selection of strains to be included in the seasonal vaccine.

Our two-scale model describes the mechanism of B cell maturation and humoral immunity in a quantitative way. In the maturation process, the B cells are first localized to one VDJ recombination and then further increase the binding affinity to the antigen by hypermutation and selection. VDJ recombinations with high initial affinity to the antigen prior to the somatic hypermutation have large chances to be selected in the maturation process and to be present in the mature B cells. The hypermutation and selection of the B cells are not deterministic, and so two zebrafish inoculated by the same type of antigen generate mature B cells with identical VDJ usages with probability p. The probability p increases with N_{size}, the number of B cells in the germinal center, decreases with n_{mut}, the hypermutation rate, and decreases with p_{cut}, the fraction of B cells surviving each round of selection. As the B cell maturation proceeds, available sequences with higher affinity to the antigen become rarer. This trend explains the rapid increase of affinity in the primary immune response and the relatively slower increase of affinity in the secondary immune response. The experimental data show that the theoretical description presented here matches aspects of the zebrafish immune system evolutionary dynamics.

We build a deterministic model equivalent to the GNK model [34], and to reproduce the observed original antigenic sin phenomenon using an ODE-based deterministic approach. Most of the parameters come from experimental data, leaving a minimal number of parameters to be estimated. The terms in the ODE system have clear physical meanings, so our model explicitly illustrates the details of the infection process.

Bibliography

[1] M. W. Deem and K. Pan, "The epitope regions of H1–subtype influenza A, with application to vaccine efficacy," *Protein Eng., Des. Sel.*, vol. 22, pp. 543–546, 2009.

[2] V. Gupta, D. J. Earl, and M. W. Deem, "Quantifying influenza vaccine efficacy and antigenic distance," *Vaccine*, vol. 24, pp. 3881–3888, 2006.

[3] A. J. Caton, G. G. Brownlee, J. W. Yewdell, and W. Gerhard, "The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype)," *Cell*, vol. 31, pp. 417–427, 1982.

[4] A. C. Shih, T. C. Hsiao, M. S. Ho, and W. H. Li, "Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution," *Proc. Natl. Acad. Sci. USA*, vol. 104, pp. 6283–6288, 2007.

[5] G. F. Weiller, "Phylogenetic profiles: A graphical method for detecting genetic recombinations in homologous sequences," *Mol Biol Evol*, vol. 15, pp. 326–335, 1998.

[6] D. L. Hartl and A. G. Clark, *Principles of population genetics.* Sunderland, Mass.: Sinauer Associates, 4th ed., 2007.

[7] J. A. Weinstein, N. Jiang, R. A. White, D. S. Fisher, and S. R. Quake, "High-throughput sequencing of the zebrafish antibody repertoire," *Science*, vol. 324,

pp. 807–810, 2009.

[8] T. Clackson and J. A. Wells, "A hot spot of binding energy in a hormone-receptor interface," *Science*, vol. 267, pp. 383–386, 1995.

[9] S. Nakajima, K. Nakajima, E. Nobusawa, J. Zhao, S. Tanaka, and K. Fukuzawa, "Comparison of epitope structures of H3HAs through protein modeling of influenza a virus hemagglutinin: Mechanism for selection of antigenic variants in the presence of a monoclonal antibody," *Microbiol Immunol*, vol. 51, pp. 1179–1187, 2007.

[10] N. Sinha, S. Mohan, C. A. Lipschultz, and S. J. Smith-Gill, "Differences in electrostatic properties at antibody-antigen binding sites: Implications for specificity and cross-reactivity," *Biophys J*, vol. 83, pp. 2946–2968, 2002.

[11] N. V. Kaverin, M. N. Matrosovich, A. S. Gambaryan, I. A. Rudneva, A. A. Shilov, N. L. Varich, N. V. Makarova, E. A. Kropotkina, and B. V. Sinitsin, "Intergenic HA–NA interactions in influenza A virus: Postreassortment substitutions of charged amino acid in the hemagglutinin of different subtypes," *Virus Res*, vol. 66, pp. 123–129, 2000.

[12] J. A. Leunissen, H. W. van den Hooven, and W. W. de Jong, "Extreme differences in charge changes during protein evolution," *J Mol Evol*, vol. 31, pp. 33–39, 1990.

[13] R. A. Sayle and E. J. Milnerwhite, "RasMol: Biomolecular graphics for all," *Trends Biochem Sci*, vol. 20, pp. 374–376, 1995.

[14] "World Health Organization Media Centre influenza fact sheet 211." `http://`

www.who.int/mediacentre/factsheets/fs211/en/index.html, accessed on August 10, 2010.

[15] "World Health Organization." http://www.who.int/csr/don/2009_05_17/en/index.html, accessed on May 25, 2009.

[16] K. Nakajima, U. Desselberger, and P. Palese, "Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950," *Nature*, vol. 274, pp. 334–339, 1978.

[17] P. Palese, T. M. Tumpey, and A. Garcia-Sastre, "What can we learn from reconstructing the extinct 1918 pandemic influenza virus?," *Immunity*, vol. 24, pp. 121–124, 2006.

[18] I. A. Wilson and N. J. Cox, "Structural basis of immune recognition of influenza-virus hemagglutinin," *Annu Rev Immunol*, vol. 8, pp. 737–771, 1990.

[19] J. J. Skehel and D. C. Wiley, "Receptor binding and membrane fusion in virus entry: The influenza hemagglutinin," *Annu Rev Biochem*, vol. 69, pp. 531–569, 2000.

[20] S. A. Frank, *Immunology and evolution of infectious disease.* Princeton, N.J.: Princeton University Press, 2002.

[21] D. P. Nayak, E. K. W. Hui, and S. Barman, "Assembly and budding of influenza virus," *Virus Res*, vol. 106, pp. 147–165, 2004.

[22] A. Portela and P. Digard, "The influenza virus nucleoprotein: A multifunctional RNA-binding protein pivotal to virus replication," *J Gen Virol*, vol. 83, pp. 723–734, 2002.

[23] M. Takeda, A. Pekosz, K. Shuck, L. H. Pinto, and R. A. Lamb, "Influenza A virus M-2 ion channel activity is essential for efficient replication in tissue culture," *J Virol*, vol. 76, pp. 1391–1399, 2002.

[24] W. S. Chen, P. A. Calvo, D. Malide, J. Gibbs, U. Schubert, I. Bacik, S. Basta, R. O'Neill, J. Schickli, P. Palese, P. Henklein, J. R. Bennink, and J. W. Yewdell, "A novel influenza A virus mitochondrial protein that induces cell death," *Nat Med*, vol. 7, pp. 1306–1312, 2001.

[25] S. D. Li, J. Y. Min, R. M. Krug, and G. C. Sen, "Binding of the influenza A virus NS1 protein to PKR mediates the inhibition of its activation by either PACT or double-stranded RNA," *Virology*, vol. 349, pp. 13–21, 2006.

[26] R. E. O'Neill, J. Talon, and P. Palese, "The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins," *EMBO J*, vol. 17, pp. 288–296, 1998.

[27] R. M. Bush, W. M. Fitch, C. A. Bender, and N. J. Cox, "Positive selection on the H3 hemagglutinin gene of human influenza virus A," *Mol. Biol. Evol.*, vol. 16, pp. 1457–1465, 1999.

[28] C. Macken, H. Lu, J. Goodman, and L. Boykin, "The value of a database in surveillance and vaccine selection," 2001. In: Osterhaus ADME, N. Cox, A. W. Hampson, editors. Options for the control of influenza IV. Elsevier; 2001, accession number ISDN38157. http://www.flu.lanl.gov/.

[29] D. J. Smith, S. Forrest, D. H. Ackley, and A. S. Perelson, "Variable efficacy of repeated annual influenza vaccination," *Proc Natl Acad Sci U S A*, vol. 96, pp. 14001–14006, 1999.

[30] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. A. M. Fouchier, "Mapping the antigenic and genetic evolution of influenza virus," *Science*, vol. 305, pp. 371–376, 2004.

[31] H. Zhou, R. Pophale, and M. W. Deem, "Computer-assisted vaccine design," *In Influenza: Molecular Virology, Horizon Scientific Press, edited by Qinghua Wang and Yizhi Jane Tao*, 2009.

[32] A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, and E. C. Holmes, "The genomic and epidemiological dynamics of human influenza A virus," *Nature*, vol. 453, pp. 615–U2, 2008.

[33] K. Pan and M. W. Deem, "Comment on Ndifon et al., "On the use of hemagglutination-inhibition for influenza surveillance: Surveillance data are predictive of influenza vaccine effectiveness"," *Vaccine*, vol. 27, pp. 5033–5034, 2009.

[34] M. W. Deem and H. Y. Lee, "Sequence space localization in the immune system response to vaccination and disease," *Phys. Rev. Lett.*, vol. 91, p. 068101, 2003.

[35] Y. Li, D. S. Carroll, S. N. Gardner, M. C. Walsh, E. A. Vitalis, and I. K. Damon, "On the origin of smallpox: Correlating variola phylogenics with historical smallpox records," *Proc Nat Acad Sci USA*, vol. 104, pp. 15787–15792, 2007.

[36] D. C. Wiley, I. A. Wilson, and J. J. Skehel, "Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation," *Nature*, vol. 289, pp. 373–378, 1981.

[37] E. Nobusawa and K. Sato, "Comparison of the mutation rates of human influenza A and B viruses," *J. Virol.*, vol. 80, pp. 3675–3678, 2006. In the amino acid level, the average mutation rate of influenza A virus is converted to 4.5×10^{-6} amino acid substitution/site/generation.

[38] J. D. Parvin, A. Moscona, W. T. Pan, J. M. Leider, and P. Palese, "Measurement of the mutation rates of animal viruses: Influenza A virus and poliovirus type 1," *J. Virol.*, vol. 59, pp. 377–383, 1986.

[39] Y. C. Liao, M. S. Lee, C. Y. Ko, and C. A. Hsiung, "Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus," *Bioinformatics*, vol. 24, pp. 505–512, 2008.

[40] A. P. Wu, Y. S. Peng, X. J. Du, Y. L. Shu, and T. J. Jiang, "Correlation of influenza virus excess mortality with antigenic variation: Application to rapid estimation of influenza mortality burden," *PLoS Comput Biol*, vol. 6, p. e1000882, 2010.

[41] Y. I. Wolf, A. Nikolskaya, J. L. Cherry, C. Viboud, E. Koonin, and D. J. Lipman, "Projection of seasonal influenza severity from sequence and serological data," *PLoS Curr*, vol. 2, p. RRN1200, 2010.

[42] N. M. Ferguson, A. P. Galvani, and R. M. Bush, "Ecological and immunological determinants of influenza evolution," *Nature*, vol. 422, pp. 428–433, 2003.

[43] C. A. Russell, T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, V. Gregory, I. D. Gust, A. W. Hampson, A. J. Hay, A. C. Hurt, J. C. de Jong, A. Kelso, A. I. Klimov, T. Kageyama, N. Komadina, A. S. Lapedes, Y. P.

Lin, A. Mosterin, M. Obuchi, T. Odagiri, A. D. M. E. Osterhaus, G. F. Rimmelzwaan, M. W. Shaw, E. Skepner, K. Stohr, M. Tashiro, R. A. M. Fouchier, and D. J. Smith, "The global circulation of seasonal influenza A (H3N2) viruses," *Science*, vol. 320, pp. 340–346, 2008.

[44] J. B. Plotkin, J. Dushoff, and S. A. Levin, "Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus," *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 6263–6268, 2002.

[45] J. J. Stewart, C. Y. Lee, S. Ibrahim, P. Watts, M. Shlomchik, M. Weigert, and S. Litwin, "A shannon entropy analysis of immunoglobulin and T cell receptor," *Mol Immunol*, vol. 34, pp. 1067–1082, 1997.

[46] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *J. Mol. Biol.*, vol. 188, pp. 415–431, 1986.

[47] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Res.*, vol. 18, pp. 6097–6100, 1990.

[48] C. Sander and R. Schneider, "Database of homology-derived protein structures and the structural meaning of sequence alignment," *Proteins*, vol. 9, pp. 56–68, 1991.

[49] P. S. Shenkin, B. Erman, and L. D. Mastrandrea, "Information-theoretical entropy as a measure of sequence variability," *Proteins*, vol. 11, pp. 297–313, 1991.

[50] M. Gerstein and R. B. Altman, "Average core structures and variability measures for protein families: Application to the immunoglobulins," *J. Mol. Biol.*, vol. 251, pp. 161–175, 1995.

[51] L. A. Mirny and E. I. Shakhnovich, "Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function," *J. Mol. Biol.*, vol. 291, pp. 177–196, 1999.

[52] K. W. Plaxco, S. Larson, I. Ruczinski, D. S. Riddle, E. C. Thayer, B. Buchwitz, A. R. Davidson, and D. Baker, "Evolutionary conservation in protein folding kinetics," *J. Mol. Biol.*, vol. 298, pp. 303–312, 2000.

[53] W. S. J. Valdar, "Scoring residue conservation," *Proteins*, vol. 48, pp. 227–241, 2002.

[54] R. M. Williamson, "Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters," *J. Theor. Biol.*, vol. 174, pp. 179–188, 1995.

[55] K. Wang and R. Samudrala, "Incorporating background frequency improves entropy-based residue conservation measures," *BMC Bioinformatics*, vol. 7, p. 385, 2006.

[56] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: Evolutionary units of three-dimensional structure," *Cell*, vol. 138, pp. 774–786, 2009.

[57] S. W. Lockless and R. Ranganathan, "Evolutionarily conserved pathways of energetic connectivity in protein families," *Science*, vol. 286, pp. 295–299, 1999.

[58] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann Math Statist*, vol. 22, pp. 79–86, 1951.

[59] E. T. Muñoz and M. W. Deem, "Epitope analysis for influenza vaccine design," *Vaccine*, vol. 23, pp. 1144–1148, 2005.

[60] J. Sun and M. W. Deem, "Statistical mechanics of the immune response to vaccines," *In Statistical Mechanics of Cellular Systems and Processes, Cambridge University Press, edited by Muhammad H. Zaman*, pp. 177–213, 2009.

[61] J. Sun, D. J. Earl, and M. W. Deem, "Localization and glassy dynamics in the immune system," *Mod Phys Lett B*, vol. 20, pp. 63–95, 2006.

[62] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345–352, 1978.

[63] S. Henikoff and J. G. Henikoff, "Amino-acid substitution matrices from protein blocks," *Proc Natl Acad Sci U S A*, vol. 89, pp. 10915–10919, 1992.

[64] J. Adachi and M. Hasegawa, "Model of amino acid substitution in proteins encoded by mitochondrial DNA," *J Mol Evol*, vol. 42, pp. 459–468, 1996.

[65] T. Müller and M. Vingron, "Modeling amino acid replacement," *J Comput Biol*, vol. 7, pp. 761–776, 2000.

[66] T. Müller, R. Spang, and M. Vingron, "Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method," *Mol Biol Evol*, vol. 19, pp. 8–13, 2002.

[67] J. L. Thorne, "Models of protein sequence evolution and their applications," *Curr Opin Genet Dev*, vol. 10, pp. 602–605, 2000.

[68] S. Veerassamy, A. Smith, and E. R. M. Tillier, "A transition probability model for amino acid substitutions from blocks," *J Comput Biol*, vol. 10, pp. 997–1010, 2003.

[69] C. Kosiol and N. Goldman, "Different versions of the Dayhoff rate matrix," *Mol Biol Evol*, vol. 22, pp. 193–199, 2005.

[70] N. Goldman and S. Whelan, "A novel use of equilibrium frequencies in models of sequence evolution," *Mol Biol Evol*, vol. 19, pp. 1821–1831, 2002.

[71] A. C. Lowen, S. Mubareka, T. M. Tumpey, A. García-Sastre, and P. Palese, "The Guinea pig as a transmission model for human influenza viruses," *Proc Natl Acad Sci U S A*, vol. 103, pp. 9988–9992, 2006.

[72] C. Janeway, P. Travers, M. Walport, and M. Shlomchik, *Immunobiology: The Immune System in Health and Disease.* New York: Garland Science, 6th ed., 2005.

[73] L. Y. H. Lee, D. L. A. Ha, C. Simmons, M. D. de Jong, N. V. V. Chau, R. Schumacher, Y. C. Peng, A. J. McMichael, J. J. Farrar, G. L. Smith, A. R. M. Townsend, B. A. Askonas, S. Rowland-Jones, and T. Dong, "Memory T cells established by seasonal human influenza A infection cross-react with avian influenza A (H5N1) in healthy individuals," *J Clin Invest*, vol. 118, pp. 3478–3490, 2008.

[74] K. Pan, K. C. Subieta, and M. W. Deem, "A novel sequence-based antigenic distance measure for H1N1, with application to vaccine effectiveness and the selection of vaccine strains," *Protein Eng., Des. Sel.*, vol. 24, pp. 291–299, 2011.

[75] R. H. Zhou, P. Das, and A. K. Royyuru, "Single mutation induced H3N2 hemagglutinin antibody neutralization: A free energy perturbation study," *J Phys Chem B*, vol. 112, pp. 15813–15820, 2008.

[76] T. Matsunaga and A. Rahman, "What brought the adaptive immune system to vertebrates? — The jaw hypothesis and the seahorse," *Immunol Rev*, vol. 166, pp. 177–186, 1998.

[77] N. S. Trede, D. M. Langenau, D. Traver, A. T. Look, and L. I. Zon, "The use of zebrafish to understand immunity," *Immunity*, vol. 20, pp. 367–379, 2004.

[78] C. Thisse and L. I. Zon, "Organogenesis—heart and wood formation from the zebrafish point of view," *Science*, vol. 295, pp. 457–462, 2002.

[79] J. A. Yoder, M. E. Nielsen, C. T. Amemiya, and G. W. Litman, "Zebrafish as an immunological model system," *Microbes Infect*, vol. 4, pp. 1469–1478, 2002.

[80] N. Danilova, V. S. Hohman, E. H. Kim, and L. A. Steiner, "Immunoglobulin variable-region diversity in the zebrafish," *Immunogenetics*, vol. 52, pp. 81–91, 2000.

[81] N. Danilova, J. Bussmann, K. Jekosch, and L. A. Steiner, "The immunoglobulin heavy-chain locus in zebrafish: Identification and expression of a previously unknown isotype, immunoglobulin Z," *Nat Immunol*, vol. 6, pp. 295–302, 2005.

[82] S. Kauffman and S. Levin, "Towards a general-theory of adaptive walks on rugged landscapes," *J Theor Biol*, vol. 128, pp. 11–45, 1987.

[83] S. A. Kauffman and E. D. Weinberger, "The *NK* model of rugged fitness landscapes and its application to maturation of the immune response," *J Theor*

Biol, vol. 141, pp. 211–245, 1989.

[84] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, "Maximum entropy models for antibody diversity," *Proc Natl Acad Sci USA*, vol. 107, pp. 5405–5410, 2010.

[85] L. D. Bogarad and M. W. Deem, "A hierarchical approach to protein molecular evolution," *Proc Natl Acad Sci USA*, vol. 96, pp. 2591–2595, 1999.

[86] J. Sun and M. W. Deem, "Spontaneous emergence of modularity in a model of evolving individuals," *Phys Rev Lett*, vol. 99, p. 228107, 2007.

[87] S. M. Anderson, A. Khalil, M. Uduman, U. Hershberg, Y. Louzoun, A. M. Haberman, S. H. Kleinstein, and M. J. Shlomchik, "Taking advantage: High-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells," *J Immunol*, vol. 183, no. 11, pp. 7314–7325, 2009.

[88] H. Y. Lee, E. Hawkins, M. S. Zand, T. Mosmann, H. L. Wu, P. D. Hodgkin, and A. S. Perelson, "Interpreting CFSE obtained division histories of B cells in vitro with Smith-Martin and cyton type models," *B Math Biol*, vol. 71, no. 7, pp. 1649–1670, 2009.

[89] S. Crotty, P. Felgner, H. Davies, J. Glidewell, L. Villarreal, and R. Ahmed, "Cutting edge: Long-term B cell memory in humans after smallpox vaccination," *J Immunol*, vol. 171, pp. 4969–4973, 2003.

[90] N. W. Schmidt, R. L. Podyminogin, N. S. Butler, V. P. Badovinac, B. J. Tucker, K. S. Bahiat, P. Lauer, A. Reyes-Sandoval, C. L. Hutchings, A. C. Moore, S. C.

Gilbert, A. V. Hill, L. C. Bartholomay, and J. T. Harty, "Memory CD8 T cell responses exceeding a large but definable threshold provide long-term immunity to malaria," *Proc Natl Acad Sci USA*, vol. 105, pp. 14017–14022, 2008.

[91] D. FitzSimons, G. Francois, A. Hall, B. McMahon, A. Meheus, A. Zanetti, B. Duval, W. Jilg, W. O. Bocher, S. N. Lu, U. Akarca, D. Lavanchy, S. Goldstein, J. Banatvala, and P. Van Damme, "Long-term efficacy of hepatitis B vaccine, booster policy, and impact of hepatitis B virus mutants," *Vaccine*, vol. 23, pp. 4158–4166, 2005.

[92] S. Brandler, M. Lucas-Hourani, A. Moris, M. P. Frenkiel, C. Combredet, M. Fevrier, H. Bedouelle, O. Schwartz, P. Despres, and F. Tangy, "Pediatric measles vaccine expressing a dengue antigen induces durable serotype-specific neutralizing antibodies to dengue virus," *PLoS Negl Trop Dis*, vol. 1, p. e96, 2007.

[93] F. S. Quan, C. Z. Huang, R. W. Compans, and S. M. Kang, "Virus-like particle vaccine induces protective immunity against homologous and heterologous strains of influenza virus," *J Virol*, vol. 81, pp. 3514–3524, 2007.

[94] K. Pan and M. W. Deem, "Predicting fixation tendencies of the H3N2 influenza virus by free energy calculation," *J. Chem. Theory Comput.*, 2011. doi: 10.1021/ct100540p.

[95] R. Durrett and V. Limic, "Rigorous results for the NK model," *Ann Probab*, vol. 31, pp. 1713–1753, 2003.

[96] C. A. Macken and A. S. Perelson, "Protein evolution on rugged landscapes," *Proc Natl Acad Sci USA*, vol. 86, pp. 6191–6195, 1989.

[97] H. Flyvbjerg and B. Lautrup, "Evolution in a rugged fitness landscape," *Phys Rev A*, vol. 46, pp. 6714–6723, 1992.

[98] P. Sibani and A. Pedersen, "Evolution dynamics in terraced NK landscapes," *Europhys Lett*, vol. 48, pp. 346–352, 1999.

[99] M. S. Lee and J. S. Chen, "Predicting antigenic variants of influenza A/H3N2 viruses," *Emerg. Infect. Dis.*, vol. 10, pp. 1385–1390, 2004.

[100] R. B. Couch, J. M. Quarles, T. R. Cate, and J. M. Zahradnik, "Clinical trials with live cold-reassortant influenza virus vaccines," in *Options for the control of influenza, UCLA Symposia on Molecular and Cellular Biology Vol. 36* (A. P. Kendal and P. A. Patriarca, eds.), pp. 223–241, New York: Alan R. Liss, 1986.

[101] W. A. Keitel, T. R. Cate, and R. B. Couch, "Efficacy of sequential annual vaccination with inactivated influenza-virus vaccine," *Am. J. Epidemiol.*, vol. 127, pp. 353–364, 1988.

[102] R. B. Couch, W. A. Keitel, T. R. Cate, J. A. Quarles, L. A. Taber, and W. P. Glezen, "Prevention of influenza virus infections by current inactivated influenza virus vaccines," in *Options for the control of influenza III* (L. E. Brown, A. W. Hampson, and R. G. Webster, eds.), pp. 97–106, Amsterdam: Elsevier Science B.V., 1996.

[103] W. A. Keitel, T. R. Cate, R. B. Couch, L. L. Huggins, and K. R. Hess, "Efficacy of repeated annual immunization with inactivated influenza virus vaccines over a five year period," *Vaccine*, vol. 15, pp. 1114–1122, 1997.

[104] K. M. Edwards, W. D. Dupont, M. K. Westrich, W. D. Plummer, P. S. Palmer, and P. F. Wright, "A randomized controlled trial of cold-adapted and inacti-

vated vaccines for the prevention of influenza A disease," *J. Infect. Dis.*, vol. 169, pp. 68–76, 1994.

[105] J. J. Treanor, K. Kotloff, R. F. Betts, R. Belshe, F. Newman, D. Iacuzio, J. Wittes, and M. Bryant, "Evaluation of trivalent, live, cold-adapted (CAIV-T) and inactivated (TIV) influenza vaccines in prevention of virus infection and illness following challenge of adults with wild-type influenza A (H1N1), A (H3N2), and B viruses," *Vaccine*, vol. 18, pp. 899–906, 1999.

[106] I. Grotto, Y. Mandel, M. S. Green, N. Varsano, M. Gdalevich, and I. Ashkenazi, "Influenza vaccine efficacy in young, healthy adults," *Clin. Infect. Dis.*, vol. 26, pp. 913–917, 1998.

[107] Z. Wang, S. Tobler, J. Roayaei, and A. Eick, "Live attenuated or inactivated influenza vaccines and medical encounters for respiratory illnesses among US military personnel," *J. Am. Med. Assoc.*, vol. 301, pp. 945–953, 2009.

[108] E. Belongia, B. Kieke, L. Coleman, J. Donahue, S. Irving, J. Meece, M. Vandermause, D. Shay, P. Gargiullo, A. Balish, A. Foust, L. Guo, S. Lindstrom, X. Xu, A. Klimov, J. Bresee, and N. Cox, "Interim within-season estimate of the effectiveness of trivalent inactivated influenza vaccine – Marshfield, Wisconsin, 2007-08 influenza season (Reprinted from vol 57, pg 393-398, 2008)," *J. Am. Med. Assoc.*, vol. 299, pp. 2381–2384, 2008.

[109] R. S. Daniels, A. R. Douglas, J. J. Skehel, and D. C. Wiley, "Antigenic and amino acid sequence analyses of influenza viruses of the H1N1 subtype isolated between 1982 and 1984," *Bull World Health Organ*, vol. 63, pp. 273–277, 1985.

[110] P. Chakraverty, P. Cunningham, G. Z. Shen, and M. S. Pereira, "Influenza in the United Kingdom 1982-85," *J. Hyg. Camb.*, vol. 97, pp. 347–358, 1986.

[111] WHO *Wkly Epidemiol Rec*, vol. 59, pp. 53–60, 1984.

[112] A. J. Hay, V. Gregory, A. R. Douglas, and Y. P. Lin, "The evolution of human influenza viruses," *Phil. Trans. R. Soc. Lond. B*, vol. 356, pp. 1861–1870, 2001.

[113] WHO *Wkly Epidemiol Rec*, vol. 61, pp. 237–244, 1986.

[114] A. P. Kendal, N. J. Cox, and M. W. Harmon, "Antigenic and genetic variation of influenza A(H1N1) viruses," pp. 119–130, 1990. In: E. Kurstak, R. G. Marusyk, F. A. Murphy, M. H. V. van Regenmortel, editors. Applied Virology Research: Virus Variability, Epidemiology and Control. Springer.

[115] I. Donatelli, L. Campitelli, A. Ruggieri, M. R. Castrucci, L. Calzoletti, and J. S. Oxford, "Concurrent antigenic analysis of recent epidemic influenza A and B viruses and quantitation of antibodies in population serosurveys in Italy," *Eur. J. Epidemiol.*, vol. 9, pp. 241–250, 1993.

[116] I. H. Brown, P. A. Harris, J. W. McCauley, and D. J. Alexander, "Multiple genetic reassortment of avian and human influenza a viruses in European pigs, resulting in the emergence of an H1N2 virus of novel genotype," *J. Gen. Virol.*, vol. 79, pp. 2947–2955, 1998.

[117] WHO *Wkly Epidemiol Rec*, vol. 67, pp. 57–64, 1992.

[118] G. F. Rimmelzwaan, J. C. de Jong, T. M. Bestebroer, A. M. van Loon, E. C. J. Claas, R. A. M. Fouchier, and A. D. M. Osterhaus, "Antigenic and genetic

characterization of swine influenza A (H1N1) viruses isolated from pneumonia patients in the Netherlands," *Virology*, vol. 282, pp. 301–306, 2001.

[119] N. Cox, A. Balish, L. Brammer, K. Fukuda, H. Hall, A. Klimov, S. Lindstrom, J. Mabry, G. Perez-Oronoz, A. Postema, M. Shaw, C. Smith, K. Subbarao, T. Wallis, and X. Xiyan, "Information for the vaccines and related biological products advisory committee, CBER, FDA. WHO collaborating center for surveillance, epidemiology and control of influenza," 2003.

[120] N. Cox, A. Balish, L. Berman, L. Blanton, L. Brammer, J. Bresee, V. Deyde, R. Donis, S. Emery, A. Foust, R. Garten, L. Gubareva, H. Hall, A. Klimov, S. Lindstrom, J. Mabry, E. Mills-Smith, A. Postema, Z. Reed, M. Shaw, B. Shu, C. Smith, S. Wang, T. Wallis, J. Winter, and X. Xiyan, "Information for the vaccines and related biological products advisory committee, CEBR, FDA. WHO collaborating center for surveillance, epidemiology and control of influenza," 2007.

[121] "WHO collaborating center for surveillance, epidemiology and control of influenza: Preliminary information for the vaccines and related biological products advisory committee, CEBR, FDA," 2008.

[122] D. M. Skowronski, G. De Serres, N. Crowcroft, N. Janjua, N. Boulianne, T. S. Hottes, and L. C. Rosella, "Seasonal influenza vaccine may be associated with increased risk of illness due to the 2009 pandemic A/H1N1 virus," *Int J Infect Dis*, vol. 14S1, pp. e321–e322, 2010.

[123] Centers for Disease Control and Prevention (CDC), "Effectiveness of 2008–09 trivalent influenza vaccine against 2009 pandemic influenza A (H1N1) — United

States, May–June 2009," *MMWR Morb Mortal Wkly Rep*, vol. 58, pp. 1241–1245, 2009.

[124] Centers for Disease Control and Prevention (CDC), "Update: Influenza activity — United States, September 28, 2008–April 4, 2009, and composition of the 2009–10 influenza vaccine," *MMWR Morb Mortal Wkly Rep*, vol. 58, pp. 369–374, 2009.

[125] Centers for Disease Control and Prevention (CDC), "Update: Influenza activity — United States, April–August 2009," *MMWR Morb Mortal Wkly Rep*, vol. 58, pp. 1009–1012, 2009.

[126] Centers for Disease Control and Prevention (CDC), "Update: Influenza activity — United States, August 30–October 31, 2009," *MMWR Morb Mortal Wkly Rep*, vol. 58, pp. 1236–1241, 2009.

[127] H. Zaraket, R. Saito, I. Sato, Y. Suzuki, D. J. Li, C. Dapat, I. Caperig-Dapat, T. Oguma, A. Sasaki, and H. Suzuki, "Molecular evolution of human influenza A viruses in a local area during eight influenza epidemics from 2000 to 2007," *Arch. Virol.*, vol. 154, pp. 285–295, 2009.

[128] P. C. Doherty and A. Kelso, "Toward a broadly protective influenza vaccine," *J Clin Invest*, vol. 118, pp. 3273–3275, 2008.

[129] H. Kelly, K. Carville, K. Grant, P. Jacoby, T. Tran, and I. Barr, "Estimation of influenza vaccine effectiveness from routine surveillance data," *PLoS One*, vol. 4, p. e5079, 2009.

[130] S. Torvaldsen and P. B. McIntyre, "Observational methods in epidemiologic

assessment of vaccine effectiveness," *Commun Dis Intell*, vol. 26, pp. 451–457, 2002.

[131] Centers for Disease Control and Prevention (CDC), "Update: Influenza activity – United States and worldwide, 1995-96 season, and composition of the 1996-97 influenza vaccine," *MMWR Morb Mortal Wkly Rep*, vol. 45, pp. 326–329, 1996.

[132] K. Pan, J. Long, H. Sun, G. J. Tobin, P. L. Nara, and M. W. Deem, "Selective pressure to increase charge in immunodominant epitopes of the H3 hemagglutinin influenza protein," *J Mol Evol*, vol. 72, pp. 90–103, 2011.

[133] WHO *Wkly Epidemiol Rec*, vol. 79, pp. 85–92, 2004.

[134] S. A. Harper, K. Fukuda, T. M. Uyeki, N. J. Cox, and C. B. Bridges, "Recommendations of the advisory committee in immunization practices.," 2004. Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report 2004;53(RR06):140.

[135] J. He and M. W. Deem, "Low–dimensional clustering reveals new influenza strains before they become dominant," *Protein Eng., Des. Sel.*, vol. 23, pp. 935–946, 2010.

[136] W. Ndifon, J. Dushoff, and S. A. Levin, "On the use of hemagglutination-inhibition for influenza surveillance: Surveillance data are predictive of influenza vaccine effectiveness," *Vaccine*, vol. 27, pp. 2447–2452, 2009.

[137] G. Meiklejohn, "Viral respiratory disease at Lowry Air Force Base in Denver, 1952–1982," *J Infect Dis*, vol. 148, pp. 775–784, 1983.

[138] H. G. Stiver, P. Graves, G. Meiklejohn, G. Schroter, and T. C. Eickhoff, "Impaired serum antibody response to inactivated influenza A and influenza B vaccine in renal transplant recipients," *Infect Immun*, vol. 16, pp. 738–741, 1977.

[139] WHO *Wkly Epidemiol Rec*, vol. 70, pp. 53–60, 1995.

[140] P. A. Gross, S. J. Sperber, A. Donabedian, S. Dran, G. Morchel, P. Cataruozolo, and G. Munk, "Paradoxical response to a novel influenza virus vaccine strain: the effect of prior immunization," *Vaccine*, vol. 17, pp. 2284–2289, 1999.

[141] WHO *Wkly Epidemiol Rec*, vol. 72, pp. 57–64, 1997.

[142] G. Winter, S. Fields, and G. G. Brownlee, "Nucleotide sequence of the hemagglutinin gene of a human influenza virus H1 subtype," *Nature*, vol. 292, pp. 72–75, 1981.

[143] E. Nobusawa, T. Aoyama, H. Kato, Y. Suzuki, Y. Tateno, and K. Nakajima, "Comparison of complete amino acid sequences and receptor binding properties among 13 serotypes of hemagglutinins of influenza A viruses," *Virology*, vol. 182, pp. 475–485, 1991.

[144] J. Felsenstein, "PHYLIP – Phylogeny Inference Package (version 3.2)," *Cladistics*, vol. 5, pp. 164–166, 1989.

[145] Y. Ina and T. Gojobori, "Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses," *Proc. Natl. Acad. Sci. USA*, vol. 91, pp. 8388–8392, 1994.

[146] T. Bedford, S. Cobey, P. Beerli, and M. Pascual, "Global migration dynamics underlie evolution and persistence of human influenza A (H3N2)," *PLoS Pathog.*, vol. 6, p. e1000918, 2010.

[147] W. M. Fitch, R. M. Bush, C. A. Bender, and N. J. Cox, "Long term trends in the evolution of H(3) HA1 human influenza type A," *Proc. Natl. Acad. Sci. USA*, vol. 94, pp. 7712–7718, 1997.

[148] J. B. Plotkin and J. Dushoff, "Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus," *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 7152–7157, 2003.

[149] V. N. Minin and M. A. Suchard, "Counting labeled transitions in continuous-time markov models of evolution," *J Math Biol*, vol. 56, pp. 391–412, 2008.

[150] Y. Y. Tseng and J. Liang, "Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: A Bayesian Monte Carlo approach," *Mol Biol Evol*, vol. 23, pp. 421–436, 2006.

[151] J. H. McDonald, "Apparent trends of amino acid gain and loss in protein evolution due to nearly neutral variation," *Mol Biol Evol*, vol. 23, pp. 240–244, 2006.

[152] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *J Comput Chem*, vol. 4, pp. 187–217, 1983.

[153] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications.* San Diego: Academic Press, 2nd ed., 2002.

[154] D. L. Beveridge and F. M. DiCapua, "Free energy via molecular simulation: Applications to chemical and biomolecular systems," *Annu Rev Biophys Biophys Chem*, vol. 18, pp. 431–492, 1989.

[155] M. Mezei and D. L. Beveridge, "Free energy simulations," *Ann NY Acad Sci*, vol. 482, pp. 1–23, 1986.

[156] A. J. Cross, "A comment on Hamiltonian parameterization in Kirkwood free energy calculations," *Ann NY Acad Sci*, vol. 482, pp. 89–90, 1986.

[157] T. C. Beutler, A. E. Mark, R. C. Vanschaik, P. R. Gerber, and W. F. Vangunsteren, "Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations," *Chem Phys Lett*, vol. 222, no. 6, pp. 529–539, 1994.

[158] M. Zacharias, T. P. Straatsma, and J. A. Mccammon, "Separation-shifted scaling, a new scaling method for lennard-jones interactions in thermodynamic integration," *J Chem Phys*, vol. 100, no. 12, pp. 9025–9031, 1994.

[159] S. Boresch and M. Karplus, "The role of bonded terms in free energy simulations: 1. theoretical analysis," *J Phys Chem A*, vol. 103, pp. 103–118, 1999.

[160] S. Boresch and M. Karplus, "The role of bonded terms in free energy simulations. 2. calculation of their influence on free energy differences of solvation," *J Phys Chem A*, vol. 103, pp. 119–136, 1999.

[161] B. Roux, "Valence selectivity of the gramicidin channel: A molecular dynamics free energy perturbation study," *Biophys J*, vol. 71, pp. 3177–3185, 1996.

[162] M. Nina, D. Beglov, and B. Roux, "Atomic radii for continuum electrostatics calculations based on molecular dynamics free energy simulations," *J Phys Chem B*, vol. 101, pp. 5239–5248, 1997.

[163] J. W. Essex, D. L. Severance, J. TiradoRives, and W. L. Jorgensen, "Monte carlo simulations for proteins: Binding affinities for trypsin–benzamidine complexes via free-energy perturbations," *J Phys Chem B*, vol. 101, pp. 9663–9669, 1997.

[164] D. J. Price and W. L. Jorgensen, "Improved convergence of binding affinities with free energy perturbation: Application to nonpeptide ligands with pp60src sh2 domain," *J Comput Aided Mol Des*, vol. 15, pp. 681–695, 2001.

[165] M. Zacharias, T. P. Straatsma, J. A. Mccammon, and F. A. Quiocho, "Inversion of receptor binding preferences by mutagenesis: Free energy thermodynamic integration studies on sugar binding to L-arabinose binding proteins," *Biochemistry*, vol. 32, pp. 7428–7434, 1993.

[166] I. Kaliman, A. Nemukhin, and S. Varfolomeev, "Free energy barriers for the N-terminal asparagine to succinimide conversion: Quantum molecular dynamics simulations for the fully solvated model," *J Chem Theory Comput*, vol. 6, pp. 184–189, 2010.

[167] A. Crespo, M. A. Marti, D. A. Estrin, and A. E. Roitberg, "Multiple-steering QM-MM calculation of the free energy profile in chorismate mutase," *J Am Chem Soc*, vol. 127, pp. 6940–6941, 2005.

[168] H. Takahashi, Y. Kawashima, T. Nitta, and N. Matubayasi, "A novel quantum mechanical/molecular mechanical approach to the free energy calculation for

isomerization of glycine in aqueous solution," *J Chem Phys*, vol. 123, p. 124504, 2005.

[169] S. L. Wang, P. Hu, and Y. K. Zhang, "Ab initio quantum mechanical/molecular mechanical molecular dynamics simulation of enzyme catalysis: The case of histone lysine methyltransferase set7/9," *J Phys Chem B*, vol. 111, pp. 3758–3764, 2007.

[170] Y. Q. Deng and B. Roux, "Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant," *J Chem Theory Comput*, vol. 2, pp. 1255–1273, 2006.

[171] D. Frenkel and A. J. C. Ladd, "New monte carlo method to compute the free energy of arbitrary solids. Application to the fcc and hcp phases of hard spheres," *J Chem Phys*, vol. 81, pp. 3188–3193, 1984.

[172] E. G. Noya, M. M. Conde, and C. Vega, "Computing the free energy of molecular solids by the einstein molecule approach: Ices XIII and XIV, hard-dumbbells and a patchy model of proteins," *J Chem Phys*, vol. 129, p. 104704, 2008.

[173] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*. San Diego: Academic Press, 2nd ed., 2002.

[174] E. J. Meijer, D. Frenkel, R. A. Lesar, and A. J. C. Ladd, "Location of melting point at 300 k of nitrogen by monte carlo simulation," *J Chem Phys*, vol. 92, pp. 7570–7575, 1990.

[175] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical-integration of cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes," *J Comput Phys*, vol. 23, pp. 327–341, 1977.

[176] C. H. Bennett, "Mass tensor molecular dynamics," *J Comput Phys*, vol. 19, pp. 267–279, 1975.

[177] R. Pomes and J. A. Mccammon, "Mass and step length optimization for the calculation of equilibrium properties by molecular dynamics simulation," *Chem Phys Lett*, vol. 166, pp. 425–428, 1990.

[178] K. A. Feenstra, B. Hess, and H. J. C. Berendsen, "Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems," *J Comput Chem*, vol. 20, no. 8, pp. 786–798, 1999.

[179] S. N. Rao, U. C. Singh, P. A. Bash, and P. A. Kollman, "Free energy perturbation calculations on binding and catalysis after mutating asn 155 in subtilisin," *Nature*, vol. 328, pp. 551–554, 1987.

[180] H. Flyvbjerg and H. G. Petersen, "Error estimates on averages of correlated data," *J Chem Phys*, vol. 91, pp. 461–466, 1989.

[181] B. R. Morgan and F. Massi, "Accurate estimates of free energy changes in charge mutations," *J Chem Theory Comput*, vol. 6, pp. 1884–1893, 2010.

[182] P. H. Hünenberger and J. A. McCammon, "Ewald artifacts in computer simulations of ionic solvation and ion-ion interaction: A continuum electrostatics study," *J Chem Phys*, vol. 110, pp. 1856–1872, 1999.

[183] F. Figueirido, G. S. Delbuono, and R. M. Levy, "On finite-size effects in computer simulations using the ewald potential," *J Chem Phys*, vol. 103, pp. 6133–6142, 1995.

[184] K. Pan and M. W. Deem, "Quantifying selection and diversity in viruses by entropy methods, with application to the hemagglutinin of H3N2 influenza," *J. R. Soc. Interface*, 2011. in press.

[185] "Ncbi influenza virus resource." `http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html`, accessed on August 10, 2010.

[186] K. Koelle, S. Cobey, B. Grenfell, and M. Pascual, "Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans," *Science*, vol. 314, pp. 1898–1903, 2006.

[187] "The FDA and CDC websites." `http://www.fda.gov/`, `http://www.cdc.gov/`, `http://www.accessdata.fda.gov/scripts/cder/drugsatfda/`.

[188] J. Petrosino, C. Cantu, and T. Palzkill, "β-lactamases: Protein evolution in real time," *Trends Microbiol*, vol. 6, pp. 323–327, 1998.

[189] H. K. Allen, L. A. Moe, J. Rodbumrer, A. Gaarder, and J. Handelsman, "Functional metagenomics reveals diverse β-lactamases in a remote Alaskan soil," *ISME J*, vol. 3, pp. 243–251, 2009.

[190] P. A. Bradford, "Extended-spectrum β-lactamases in the 21st century: Characterization, epidemiology, and detection of this important resistance threat," *Clin Microbiol Rev*, vol. 14, pp. 933–951, 2001.

[191] K. Bush, "Extended-spectrum β-lactamases in North America, 1987-2006," *Clin Microbiol Infect*, vol. 14, pp. 134–143, 2008.

[192] F. Perez, A. Endimiani, K. M. Hujer, and R. A. Bonomo, "The continuing challenge of ESBLs," *Curr Opin Pharmacol*, vol. 7, pp. 459–469, 2007.

[193] "The Lahey Website on extended spectrum resistance bacteria. The Lahey Website has extended spectrum clinical isolates, which had been under intense antibiotic selection at active sites.." `http://www.lahey.org/Studies/`.

[194] J. Chaves, M. G. Ladona, C. Segura, A. Coira, R. Reig, and C. Ampurdanes, "SHV-1 β-lactamase is mainly a chromosomally encoded species-specific enzyme in *Klebsiella pneumoniae*," *Antimicrob Agents Chemother*, vol. 45, pp. 2856–2861, 2001.

[195] R. S. Harris, S. Longerich, and S. M. Rosenberg, "Recombination in adaptive mutation," *Science*, vol. 264, pp. 258–260, 1994.

[196] R. G. Ponder, N. C. Fonville, and S. M. Rosenberg, "A switch from high-fidelity to error-prone DNA double-strand break repair underlies stress-induced mutation," *Mol Cell*, vol. 19, pp. 791–804, 2005.

[197] M. Barlow, J. Fatollahi, and M. Salverda, "Evidence for recombination among the alleles encoding TEM and SHV β-lactamases," *J Antimicrob Chemother*, vol. 63, pp. 256–259, 2009.

[198] D. M. Livermore, "Has the era of untreatable infections arrived?," *J Antimicrob Chemother*, vol. 64, pp. 29–36, 2009.

[199] D. M. Livermore and N. Woodford, "The β-lactamase threat in *Enterobacteriaceae*, *Pseudomonas* and *Acinetobacter*," *Trends Microbiol*, vol. 14, pp. 413–420, 2006.

[200] S. Bailey, "The CCP4 suite: Programs for protein crystallography," *Acta Crystallogr D Biol Crystallogr*, vol. 50, pp. 760–763, 1994.

[201] B. Wiedemann, C. Kliebe, and M. Kresken, "The epidemiology of β-lactamases," *J Antimicrob Chemother*, vol. 24, pp. 1–22, 1989.

[202] P. M. Bennett, "Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria," *Br J Pharmacol*, vol. 153, pp. S347–S357, 2008.

[203] A. Carattoli, "Resistance plasmid families in *Enterobacteriaceae*," *Antimicrob Agents Chemother*, vol. 53, pp. 2227–2238, 2009.

[204] D. J. Earl and M. W. Deem, "Evolvability is a selectable trait," *Proc Natl Acad Sci USA*, vol. 101, pp. 11531–11536, 2004.

[205] W. P. C. Stemmer, "Rapid evolution of a protein *in vitro* by DNA shuffling," *Nature*, vol. 370, pp. 389–391, 1994.

[206] H. Knothe, P. Shah, V. Krcmery, M. Antal, and S. Mitsuhashi, "Transferable resistance to cefotaxime, cefoxitin, cefamandole and cefuroxime in clinical isolates of Klebsiella pneumoniae and Serratia marcescens," *Infection*, vol. 11, pp. 315–317, 1983.

[207] A. Baraniak, J. Fiett, A. Mrowka, J. Walory, W. Hryniewicz, and M. Gniadkowski, "Evolution of TEM-type extended-spectrum β-lactamases in clinical *Enterobacteriaceae* strains in poland," *Antimicrob Agents Chemother*, vol. 49, pp. 1872–1880, 2005.

[208] F. C. Tenover, "Mechanisms of antimicrobial resistance in bacteria," *Am J Med*, vol. 119, pp. S3–S10, 2006.

[209] E. J. Feil, M. C. Enright, and B. G. Spratt, "Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between neisseria meningitidis and streptococcus pneumoniae," *Res Microbiol*, vol. 151, pp. 465–469, 2000.

[210] J. Sun, D. J. Earl, and M. W. Deem, "Glassy dynamics in the adaptive immune response prevents autoimmune disease," *Phys Rev Lett*, vol. 95, p. 148104, 2005.

[211] "University of manchester bioinformatics website was the main source of programs and their references and documentation.." `http://www.bioinf.manchester.ac.uk/recombination/programs.shtml`.

[212] `http://www.ub.es/dnasp/`.

[213] R. R. Hudson and N. L. Kaplan, "Statistical properties of the number of recombination events in the history of a sample of DNA sequences," *Genetics*, vol. 111, pp. 147–164, 1985.

[214] R. R. Hudson, "Estimating the recombination parameter of a finite population model without selection," *Genet Res*, vol. 50, pp. 245–250, 1987.

[215] G. McVean, P. Awadalla, and P. Fearnhead, "A coalescent-based method for detecting and estimating recombination from gene sequences," *Genetics*, vol. 160, pp. 1231–1241, 2002.

[216] G. A. T. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly, "The fine-scale structure of recombination rate variation in the human genome," *Science*, vol. 304, pp. 581–584, 2004.

[217] S. Suerbaum, J. M. Smith, K. Bapumia, G. Morelli, N. H. Smith, E. Kunstmann, I. Dyrek, and M. Achtman, "Free recombination within *Helicobacter pylori*," *Proc Natl Acad Sci U S A*, vol. 95, pp. 12619–12624, 1998.

[218] I. B. Jakobsen and S. Easteal, "A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences," *Comput Appl Biosci*, vol. 12, pp. 291–295, 1996.

[219] `http://pubmlst.org/software/analysis/start2/`, `http://pubmlst.org/software/analysis/start/manual/`.

[220] J. M. Smith, "Analyzing the mosaic structure of genes," *J Mol Evol*, vol. 34, pp. 126–129, 1992.

[221] S. Sawyer, "Statistical tests for detecting gene conversion," *Mol Biol Evol*, vol. 6, pp. 526–538, 1989.

[222] T. C. Bruen, H. Philippe, and D. Bryant, "A simple and robust statistical test for detecting the presence of recombination," *Genetics*, vol. 172, pp. 2665–2681, 2006.

[223] D. H. Huson, "Splitstree: analyzing and visualizing evolutionary data," *Bioinformatics*, vol. 14, pp. 68–73, 1998.

[224] G. Drouin, F. Prat, M. Ell, and G. D. P. Clarke, "Detecting and characterizing gene conversions between multigene family members," *Mol Biol Evol*, vol. 16, pp. 1369–1390, 1999.

[225] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Mol Biol Evol*, vol. 23, pp. 254–267, 2006. Software available

from `www.splitstree.org`.

[226] D. A. Drummond, J. J. Silberg, M. M. Meyer, C. O. Wilke, and F. H. Arnold, "On the conservative nature of intragenic recombination," *Proc Natl Acad Sci U S A*, vol. 102, pp. 5380–5385, 2005.

[227] I. Hellmann, I. Ebersberger, S. E. Ptak, S. Paabo, and M. Przeworski, "A neutral explanation for the correlation of diversity with recombination rates in humans," *Am J Hum Genet*, vol. 72, pp. 1527–1535, 2003.

[228] W. P. Hanage, C. Fraser, J. Tang, T. R. Connor, and J. Corander, "Hyper-recombination, diversity, and antibiotic resistance in Pneumococcus," *Science*, vol. 324, pp. 1454–1457, 2009.

[229] C. Mabilat and P. Courvalin, "Development of "oligotyping" for characterization and molecular epidemiology of TEM β-lactamases in members of the family *Enterobacteriaceae*," *Antimicrob Agents Chemother*, vol. 34, pp. 2210–2216, 1990.

[230] K. Vetsigian and N. Goldenfeld, "Global divergence of microbial genome sequences mediated by propagating fronts," *Proc Natl Acad Sci U S A*, vol. 102, pp. 7332–7337, 2005.

[231] D. M. Weinreich, N. F. Delaney, M. A. DePristo, and D. L. Hartl, "Darwinian evolution can follow only very few mutational paths to fitter proteins," *Science*, vol. 312, pp. 111–114, 2006.

[232] J. E. Mroczkowska and M. Barlow, "Fitness trade-offs in bla_{TEM} evolution," *Antimicrob Agents Chemother*, vol. 52, pp. 2340–2345, 2008.

[233] C. Berek, A. Berger, and M. Apel, "Maturation of the immune response in germinal centers," *Cell*, vol. 67, pp. 1121–1129, 1991.

[234] J. M. Park and M. W. Deem, "Correlations in the T-cell response to altered peptide ligands," *Physica A*, vol. 341, pp. 455–470, 2004.

[235] E. Muñoz and M. W. Deem, "Amino acid alphabet size in protein evolution experiments: better to search a small library thoroughly or a large library sparsely?," *Protein Eng Des Sel*, vol. 21, pp. 311–317, 2008.

[236] J. Jacob, R. Kassir, and G. Kelsoe, "In situ studies of the primary immune-response to (4-hydroxy-3-nitrophenyl)acetyl. I. The architecture and dynamics of responding cell populations," *J Exp Med*, vol. 173, pp. 1165–1175, 1991.

[237] K. G. C. Smith, A. Light, G. J. V. Nossal, and D. M. Tarlinton, "The extent of affinity maturation differs between the memory and antibody-forming cell compartments in the primary immune response," *EMBO J*, vol. 16, pp. 2996–3006, 1997.

[238] J. S. Zhang and E. I. Shakhnovich, "Optimality of mutation and selection in germinal centers," *PLoS Comput Biol*, vol. 6, p. e1000800, 2010.

[239] G. I. Marchuk, R. V. Petrov, A. A. Romanyukha, and G. A. Bocharov, "Mathematical model of antiviral immune response. I. Data analysis, generalized picture construction and parameters evaluation for hepatitis B," *J Theor Biol*, vol. 151, pp. 1–40, 1991.

[240] G. A. Bocharov and A. A. Romanyukha, "Mathematical model of antiviral immune response III. Influenza A virus infection," *J Theor Biol*, vol. 167, pp. 323–360, 1994.

[241] G. T. Belz, D. Wodarz, G. Diaz, M. A. Nowak, and P. C. Doherty, "Compromised influenza virus-specific CD8(+)-T-cell memory in CD4(+)-T-cell-deficient mice," *J Virol*, vol. 76, pp. 12388–12393, 2002.

[242] P. Baccam, C. Beauchemin, C. A. Macken, F. G. Hayden, and A. S. Perelson, "Kinetics of influenza A virus infection in humans," *J Virol*, vol. 80, pp. 7590–7599, 2006.

[243] D. B. Chang and C. S. Young, "Simple scaling laws for influenza A rise time, duration, and severity," *J Theor Biol*, vol. 246, pp. 621–635, 2007.

[244] B. Hancioglu, D. Swigon, and G. Clermont, "A dynamical model of human immune response to influenza A virus infection," *J Theor Biol*, vol. 246, pp. 70–86, 2007.

[245] M. A. Nowak and R. M. May, *Virus dynamics: Mathematical principles of immunology and virology.* Oxford, New York: Oxford University Press, 2000.

[246] J. L. Segovia-Juarez, S. Ganguli, and D. Kirschner, "Identifying control mechanisms of granuloma formation during M-tuberculosis infection using an agent-based model," *J Theor Biol*, vol. 231, pp. 357–376, 2004.

[247] C. Beauchemin, J. Samuel, and J. Tuszynski, "A simple cellular automaton model for influenza A viral infections," *J Theor Biol*, vol. 232, pp. 223–234, 2005.

[248] R. Flindt and N. Solomon, *Amazing numbers in biology.* Berlin, New York: Springer-Verlag, 2006.

[249] K. P. Keenan, J. W. Combs, and E. M. Mcdowell, "Regeneration of hamster tracheal epithelium after mechanical injury. I. Focal lesions: Quantitative morphologic study of cell proliferation," *Virchows Arch B Cell Pathol Incl Mol Pathol*, vol. 41, pp. 193–214, 1982.